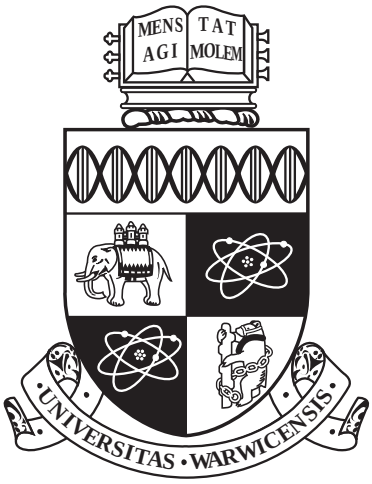


The Temporal and Spatial Analysis of Single Cell Gene Expression

by

Kirsty Hey

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics, University of Warwick

March 2015

THE UNIVERSITY OF
WARWICK

For Mark

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor Dr Bärbel Finkenstädt for her guidance and support throughout the past years. Her advice and encouragement have helped shape both this research and me as a researcher. In addition, I would like to thank Prof. David Rand and the research group as a whole (both current and past members) for many stimulating discussions and general solidarity at Friday morning group meetings. The collaborative nature of this project has been truly enjoyable due to the general enthusiasm to share both data and ideas. Acknowledgement should be given to the labs of both Mike White and Julian Davis. In particular, I'd like to thank Karen Featherstone for her patience with my many biology questions and also for letting me see first hand the experimental procedures.

The Department of Statistics has provided me with many opportunities to broaden my knowledge and expertise through funding to attend numerous conferences and workshops. Additionally, the Engineering and Physical Sciences Research Council has provided me with financial support (EPSRC grant number ASTAA1112.KXH) throughout my studies.

To my friends, old and new, whether it's Axel for patiently describing SMC, Kasia for introducing me to green tea, Silvia for shared commiserations of non-converging MCMC or Lorna and Cata for the shared enjoyment of apple cake. Along with many others you have made my time at Warwick all the more enjoyable.

A special thanks to my family, for always being there. Words can't express the gratitude I feel for all the love, kindness and support you've given me. Thank you.

DECLARATIONS

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

All experimental data discussed in this thesis have been kindly provided by the labs of Mike White and Julian Davis at the University of Manchester. In particular, acknowledgement should be given to Karen Featherstone who performed the experimental procedures.

In addition, Chapter 7 is collaborative work with Hiroshi Momiji whose contribution greatly aided both the methodology and analysis.

Parts of this thesis have been published by the author in the following publication:

***K Hey**, H Momiji, K Featherstone, J Davis, M White, D Rand and B Finkenstädt* (2015). **Inference for a Transcriptional Stochastic Switch Model from Single Cell Imaging Data**. *Biostatistics*. DOI:10.1093/biostatistics.kxv010.

ABSTRACT

It is the aim of this thesis to provide a rigorous and comprehensive analysis of single cell gene expression data. Specifically, the focus is on expression of the human Prolactin gene, which can be measured in intact tissue samples via a reporter process. To do this, we develop a robust statistical procedure, the stochastic switch model, to model the transcriptional regulation within single cells that properly accounts for both intrinsic and extrinsic variability whilst also incorporating a realistic measurement process. The stochastic switch model provides a highly flexible framework for coupling the regulatory system without the need for detailed prior knowledge of the underlying regulatory mechanisms. In this thesis, this methodology is applied to numerous datasets to find different regulatory behaviour evident in different biological conditions. Moreover, since the data provided has in addition a representative spatial resolution, we investigate how the spatial organisation of the expressing cells changes in these different biological conditions. This is achieved via spatial point processes and makes use of the recently developed hybrid Gibbs processes. The thesis ends by revisiting the transcriptional regulation within single cells and how analysing these processes in space reveals evidence of cell signalling. From this evidence, various semi-mechanistic models are derived with attention focused on model identifiability. Consequently, this thesis provides both methods and analysis for the temporal, the spatial and the spatio-temporal data regarding single cell gene expression.

CHAPTER 1

INTRODUCTION

So perhaps the best thing to do is to stop writing introductions and get on with the book.

A.A. Milne, Winnie-the-Pooh

During the last few decades there have been huge advances in experimental biology and the techniques employed that have yielded many interesting and important insights into the underlying biological processes. With increasing frequency, many of these processes have been shown to have stochastic variation at the molecular level. Moreover, as experiments become increasingly complex, relating the biology to the measured outcomes becomes challenging. It has therefore become clear that robust and rigorous statistical analysis is key to extracting information from such experiments. The objective of this research is to provide robust statistical methodology for the analysis of gene expression dynamics where observations consist of single cell measurements of reporter protein levels. Section 1.1 provides a basic introduction to the fundamentals of gene expression and the particular experimental techniques to which we focus our attention. Motivating data is presented in Section 1.2 and provides the framework of this thesis, outlined in Section 1.3.

1.1 Biological Background

1.1.1 Gene Expression

Information encoded within individual genes is transferred through the process of gene expression. This process results in the synthesis of some gene product, which most commonly is a specific protein that is required for some role within the body. Examples include the production of hormones that in turn regulate physiology and behaviour. Consequently, the regulation of gene expression within single cells enables individual cells to control and respond to the molecular environment. For example, the expression of the Prolactin hormone will be highly up-regulated in response to lactation in order to produce increasing amounts of milk.

Gene expression typically consists of three main processes:

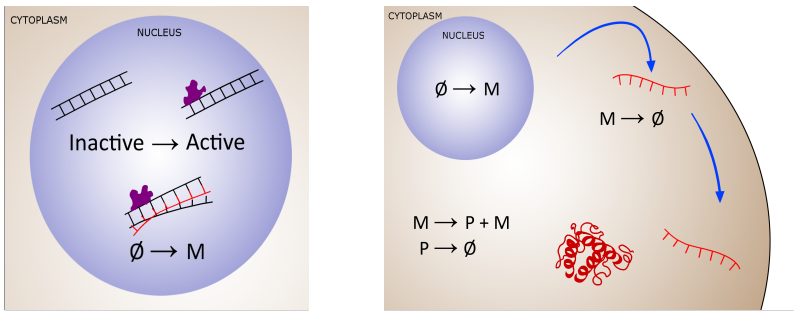
Gene activation/regulation. For a gene to be expressed it first has to be activated. This activation occurs through a series of reactions including the binding and/or unbinding of transcription factors to the gene promoter region of the DNA.

Transcription. Once a gene has been activated, it can be transcribed to result in the production of mRNA molecules.

Translation. The mRNA molecules will then move out of the cell nucleus and into the cytoplasm to either be degraded or translated into proteins. Proteins are the final product of gene expression and will move on to play some role either within the same cell or somewhere else within the body.

The above processes are depicted in Figure 1.1, which provides a coarse level representation. There are many other molecular processes involved in gene expression, including protein folding and unfolding, which we do not consider in our applications due to the limited resolution of the data available.

In single cells, gene expression is made up of fundamentally stochastic processes (Raj and Van Oudenaarden, 2008) due to both *intrinsic* and *extrinsic* variation. Intrinsic variability is the variation observed between the molecular processes of identical gene copies, which arise from random microscopic events determining these processes. For instance the reactions involved within these processes will occur randomly according to the law of mass action and thus the event of a successful reaction will be a random variable with probability depending on the reaction rate. In comparison,



(a) (b)

Figure 1.1: Illustration of gene expression where in a) the gene is activated through the binding of certain molecules to allow mRNA molecules to be created through transcription. Following this, shown in b), the mRNA molecules then move out of the nucleus and into the cytoplasm to either degrade or be translated into proteins.

extrinsic variability is the intercellular variability of gene expression caused by fluctuations in molecular activity due to both interacting processes and randomness in molecular machinery (Elowitz et al., 2002). This can be interpreted mathematically by considering fluctuations in the reaction rate constants across differing cells. Consequently, one must incorporate both these sources of stochasticity when analysing gene expression within single cells.

Analysing gene expression data has attracted much attention (see for example, Kærn et al. (2005); Raj and Van Oudenaarden (2008); Spiller et al. (2010); Zechner et al. (2014)) with many interesting features having been elicited. The main feature we investigate in this thesis is the pulsatile behaviour of expression that has been found for many different genes (Suter et al., 2011; Harper et al., 2011). In particular, this pulsatility has been shown to be highly variable between individual cells. Recent papers (Blake et al., 2006; Paszek et al., 2010; Harper et al., 2010, 2011) hypothesise this varying pulsatile behaviour to be necessary for robust tissue level response to a range of physiological conditions and we will investigate this further for a single gene application.

1.1.2 Experimental Techniques

There are many different techniques used to collect data on gene regulation and expression. Examples include microarray experiments where the concentration of mRNA molecules is measured from aggregated data or qPCR (quantitative polymerase chain reaction) experiments, which measure something proportional to the concentration of mRNA from a specific target gene. In particular, these data are

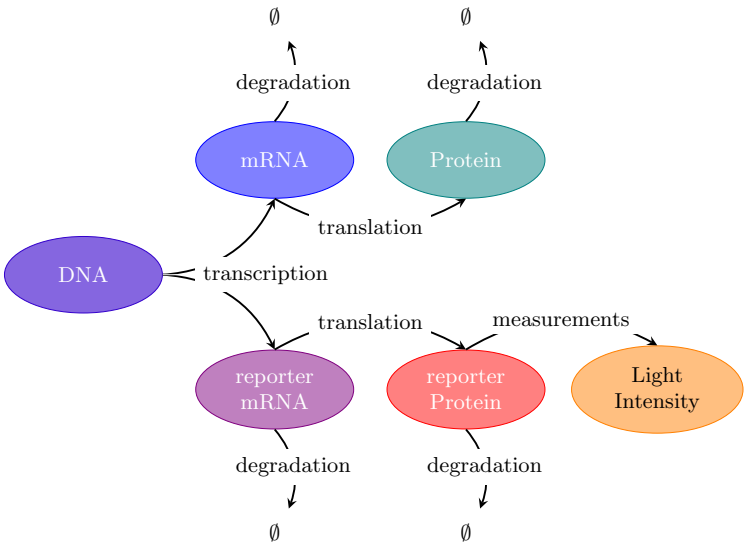


Figure 1.2: A diagrammatic representation of how the measurement process through reporter genes relates to the underlying process of native gene expression. Specifically, both native and reporter mRNA are transcribed in parallel, after which the two processes become independent with the reporter mRNA being translated into reporter protein independently of the translation of native mRNA. The reporter protein can then be measured through light microscopy techniques.

aggregated over many cells and represent an average behaviour. Here, however, we are particularly interested in quantifying the dynamics within individual cells. Measurements on individual cells can be obtained through live cell imaging techniques and has proven successful for studying the stochastic temporal expression dynamics of reporter genes (Stephens and Allan, 2003; Spiller et al., 2010). A reporter gene is a gene that is inserted into cell DNA and engineered to be controlled by the same promoter as the gene of interest. Examples of reporter genes are the genes coding for fluorescent and luminescent proteins, which can be measured indirectly through light microscopy. Figure 1.2 gives a diagrammatic representation of how expression of the reporter gene relates to native gene expression. Specifically, since the reporter is under the control of the native gene promoter, both reporter and native mRNA will be transcribed in parallel. The reporter mRNA will then be translated into reporter protein independently of the native mRNA. Levels of these reporter proteins can then be measured indirectly through light microscopy. For instance for a green fluorescent protein, under a laser, levels of fluorescence can be measured and will be proportional (with error) to the number of reporter protein molecules. Consequently, reporter gene expression is summarised by the reactions (Elowitz et al.,

2002; Nelson et al., 2004; Paulsson, 2005),

$$\text{DNA} \xrightarrow{\beta(t)} \text{mRNA}, \qquad \text{mRNA} \xrightarrow{\delta_m} \emptyset, \tag{1.1}$$

$$\text{mRNA} \xrightarrow{\alpha} \text{mRNA} + \text{Protein}, \quad \text{Protein} \xrightarrow{\delta_p} \emptyset, \tag{1.2}$$

where the superscript for each reaction denotes the corresponding reaction rate. The rate of transcription, $\beta(t)$, will be time varying as it depends on the activation state of the underlying gene and will therefore represent the regulation of the underlying gene promoter. Moreover, these transcriptional dynamics of the reporter will relate to the transcriptional dynamics of the native gene (Finkenstädt et al., 2008; Harper et al., 2011) due to the coupling depicted in Figure 1.2. Consequently, it is these transcriptional dynamics that relate the observed reporter gene expression dynamics to the regulation of the native gene of interest, also termed the target gene.

1.2 Motivating Data

Representative data following the above reporter construction are shown in Figure 1.3. The target gene for these data is the Prolactin (PRL) gene whose regulation is of particular interest, due to its important roles in mammalian reproduction and also its frequent over-production by pituitary adenomas (Harper et al., 2010). In order to obtain measurements, a line of transgenic rats was created. Specifically, the rat genome was modified so that the native DNA contained the human Prolactin promoter that in turn controlled the expression of a destabilised green fluorescent reporter protein. Since rats naturally produce the regulatory transcription factors that bind to the human Prolactin promoter, this rat model allows one to study the regulation of the human Prolactin gene within mammalian tissue, specifically rat pituitary tissue. A number of different samples have been collected from different physiological conditions. Namely, pituitary tissue has been taken from animals in different stages of development ranging from adulthood to post-natal day 1.5 and embryonic day 18.5. Further details of the reporter construct used and associated experimental framework can be found in Semprini et al. (2009); Harper et al. (2010) and Featherstone et al. (2011).

As an example, two representative datasets are shown in Figure 1.3 where the top row corresponds to a pituitary tissue sample taken from an adult male rat and the bottom row corresponding to a pituitary tissue sample of a post-natal day 1.5 rat. Figures 1.3 a) and e) show the raw image file from which an area of approximately

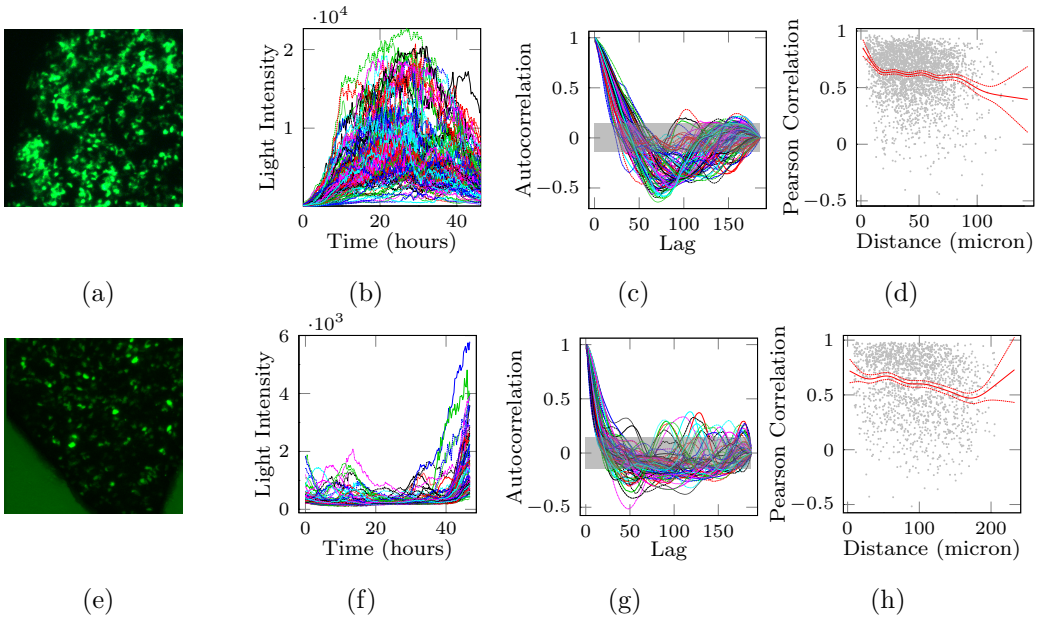


Figure 1.3: Example data of a green fluorescent reporter under the control of the human Prolactin promoter. The top row shows data obtained from an adult male rat pituitary tissue sample with the bottom row corresponding to data obtained from a post-natal day 1.5 (P1.5) rat pituitary. Figures a) and e) show the raw image file from which an area of approximately 100 cells were tracked every 15 minutes over a 46 hour period to obtain the time series data in Figures b) and f). The corresponding autocorrelation plots are shown in Figures c) and g). Figures d) and h) then show the relationship between the Pearson correlation coefficient between each pair of time series and their Euclidean distance. The fitted red line is a penalised regression spline with associated 95% confidence bands about the mean response.

100 cells were tracked every 15 minutes over a 46 hour period to obtain the time series data in Figures 1.3 b) and f). These data exemplify some of the issues arising in single cell imaging data. In particular, the data consist of time discrete observations of the underlying reporter protein levels measured indirectly through the imaging process. Moreover, reporter mRNA abundance is completely unobserved. There is clear stochasticity in the data and in addition, the corresponding autocorrelation functions in Figures 1.3 c) and g) demonstrate evidence of clear pulsatile behaviour as has also been observed in Harper et al. (2011).

1.2.1 Temporal Dynamics

Little is known about the regulation mechanisms driving Prolactin gene expression. Previous papers (Harper et al., 2010) have analysed similar data to that shown in Figure 1.3 by grouping the time series data into clusters of similarly behaved

patterns. The authors suggest several hypotheses for these differing patterns including structural features of the pituitary, functional speciation of lactotroph cells (Christian et al., 2007) or local paracrine signalling. However, the authors did not take into account the stochastic nature of the data nor did they relate the observed dynamics back to the native process. Consequently, this motivates Part I of this thesis, to provide a methodology that analyses the temporal dynamics of reporter gene expression data and relates back to the native process regulating the native gene.

Since it is only through transcription that the reporter and native processes are coupled, the main aim is to back calculate from light intensity measurements back to reporter protein levels, back to reporter mRNA and finally to the underlying transcriptional dynamics. As specified before, these transcriptional dynamics model the regulation of the underlying gene. The most commonly assumed model for gene regulation is the binary switch model (Peccoud and Ycart, 1995; Kærn et al., 2005; Larson et al., 2009; Suter et al., 2011; Harper et al., 2011) where,

$$\beta(t) = \begin{cases} \beta_0 & \text{if } t \text{ is in an on-phase} \\ \beta_1 & \text{if } t \text{ is in an off-phase.} \end{cases}$$

Here transcription can take only one of two values corresponding to the gene being in an active or inactive state. Typically it is assumed that β_0 is low or possibly zero for an inactive gene state. This simple model has been used extensively to infer dynamics of gene regulation for various systems. For example, both Suter et al. (2011) and Harper et al. (2011) found evidence of a refractory period where a gene was regulated by a three state Markov process switching between on, off and refractory states with associated transcriptional states given by,

$$\beta(t) = \begin{cases} \beta_0 & \text{if } t \text{ is in an on-phase} \\ \beta_1 & \text{if } t \text{ is in either an off- or refractory- phase.} \end{cases}$$

More complex structures for gene regulation can also be described by this binary model, for example, Sanchez et al. (2013) produce the predicted transition times of the binary switch given a four state promoter, with transcription occurring in only a single promoter state. In addition to reconstructing gene expression dynamics, the binary switch model has also been used to infer transcription factor interactions (Sanguinetti et al., 2009; Opper and Sanguinetti, 2010).

Other transcriptional functions used in the literature include a sinusoidal func-

tion (Komorowski et al., 2010) motivated by circadian clock data (Chabot et al., 2007), or an exponential function relating to the presence of experimental stimulus (Finkenstädt et al., 2008). Hill functions have also been used extensively (see for example Rosenfeld et al. (2005)) to model transcription with additional dependence on the levels of activator or repressor that are present in the cell nucleus.

In contrast, Jenkins et al. (2013) extended the binary switch model to allow for multiple rates of transcription, where,

$$\beta(t) = \beta_i \text{ for } t \in [s_{i-1}, s_i). \tag{1.3}$$

Here the transcriptional rate of a gene is piecewise constant over intervals with changes in transcriptional regulation at the unknown switch times $s_1, ..., s_K$, associated with unobserved transcriptional events. Although the binary switch has a simple biological interpretation, restricting transcription to binary states may not capture the full range of cellular activity as other events such as limiting/competing transcription factors may influence gene regulation. Jenkins et al. (2013) fitted the multiple state switch model to aggregated mRNA populations as observed in microarray analyses and found that the approach is general enough to describe a wide range of observed dynamic patterns in gene expression including oscillatory behaviour with asymmetric cycles of varying amplitude.

In what follows we will demonstrate how the multi-state switch model may be used to provide possible hypotheses for the mechanisms of gene regulation including (up to limited identifiability) interacting transcription factors. Consequently, embedding the multi-state switch model within a stochastic framework will form the basis of Part I of this thesis.

1.2.2 Spatial Organisation

One of the unique aspects of these data is the spatial resolution, examples of which are shown in Figure 1.3 a) for an adult male tissue and 1.3 e) for a post-natal day 1.5 tissue. Typically, these imaging experiments are performed on dispersed cells within a culture, however, here the data consist of cells within intact tissue to provide a spatial resolution representative of the spatial organisation within a native system.

Consequently, these data provide a unique opportunity to also study the spatial organisation of Prolactin producing cells (called lactotrophs). In addition, we have a number of datasets obtained from animals in different stages of development and

it is therefore of interest to study the changes (if any) in spatial organisation of cell location to gain insight into the tissue architecture. For instance, the positioning of lactotrophs may be quantitatively different during different stages of development.

Ideally one would wish to model the changing spatial organisation mechanistically. For example spring-bead models provide a highly flexible framework for modelling spatial dynamics. Applications include modelling of plant tissue incorporating cell dynamics such as growth and division (Shapiro and Mjolsness, 2001; Shapiro et al., 2011), or applications in dissipative particle physics (Groot and Warren, 1997) such as polymer structures (Symeonidis et al., 2005) and platelet aggregation in arteries (Pivkin et al., 2009). Lattice based methods have also been used in similar applications, for instance Dobrescu and Purcarea (2009) applied a cellular automata model to describe tumour growth.

However, the data we have available substantially limits the feasibility of inference for such mechanistic approaches. In particular, although we have replicate tissue samples from different stages of development, each sample comes from independent tissues and consequently there is no mechanistic way of linking the data through developmental stage. Therefore a statistical approach becomes much more appropriate and in particular we will use spatial point processes to provide a statistical description of tissue architecture at different stages of development. This will form the basis of Part II of this thesis.

1.2.3 Spatial Coordination

It is a natural extension, given the motivating data consist of both temporal and spatial resolutions, to consider the spatio-temporal relationship. In particular, Figures 1.3 d) and h) show the relationship between the Pearson correlation coefficient between any pair of time series and their Euclidean distance. It is clear that within the adult tissue, there is an increased synchronicity at shorter distances perhaps indicative of a cell signalling mechanism. However, this synchronicity is not evident in the P1.5 tissue and motivates further investigation. In particular, it is desirable to incorporate any spatio-temporal coupling at the native gene level rather than at the reporter protein level. Therefore the spatio-temporal modelling should be based on the back calculated transcriptional profiles and will be presented in Part III of the thesis.

1.3 Thesis Outline

As motivated by the data presented in the previous section this thesis is separated into the three following parts.

Part I

Part I provides a rigorous statistical methodology for estimation of the transcriptional dynamics of single cell gene expression as described in Hey et al. (2015). In particular, it is the aim of this work to embed the multiple state switch model within a general stochastic modelling and inference framework, namely the *stochastic switch model* (SSM), to study gene expression dynamics at the single cell level. Our approach is derived on the basis of a stochastic reaction network (SRN) to capture the intrinsic variability whilst also introducing a realistic measurement equation with unknown parameters in order to fit the model to experimental single cell imaging time series accounting for variability due to the measurement process. The resulting model provides an approach that is both scientifically interpretable and flexible enough to capture a wide range of stochastic dynamics observed in longitudinal single cell imaging data including irregular pulsatile behaviour.

Chapter 2 introduces stochastic reaction networks and their associated approximations. In addition to the approximations common within the literature, we derive a model specific approximation termed the birth-death approximation or BDA. We find that a state space representation provides a unifying framework for stochastic reaction networks in the presence of a measurement process. Chapter 3 consists of a discussion of the inferential techniques for these state space models along with an extensive simulation study. A complete application to single cell imaging data is presented in Chapter 4 along with extensive analyses of the posterior transcriptional profiles.

Part II

Part II consists of a comprehensive analysis of the spatial organisation of lactotroph cells during development of the mammalian pituitary. This is achieved within a spatial point process framework and allows us to identify many interesting features of the data. We find that the recently developed family of hybrid Gibbs point processes provides a flexible framework that can encompass the different features

identified at different developmental stages. In particular, as tissues mature from post natal day 1.5 to adulthood, we find evidence of the development of spatial clustering, which may result in more cohesive networks. This is investigated in some detail. Chapter 5 provides a detailed introduction to spatial point processes with their application to data presented in Chapter 6.

Part III

Part III presents the groundwork for modelling temporal transcriptional dynamics within single cells whilst incorporating a spatial resolution. Simulation models are constructed based on the posterior analysis of the multi-state switch model applied to single cell imaging data. In particular, we construct a spatial transcriptional model so that the marginal temporal dynamics inferred from Part I are recovered when integrating over the spatial domain. through the analysis of real data, several spatio-temporal dependencies are observed in the posterior transcriptional profiles. Interestingly, we see that certain spatio-temporal features can be recovered by a completely separable space-time model where spatial location is given by the analysis of Part II and are therefore not captured through signalling mechanisms. In contrast, other space-time features can only be reproduced through a signalling component through an explicit spatial dependence in the transcriptional profile. We therefore investigate different formulations that can be associated to different signalling mechanisms. Attention is focussed on model identifiability where we investigate the implications of different data resolutions to provide recommendations to ensure optimal identifiability.

A complete exploratory analysis of the spatial distribution of the posterior transcriptional profiles is presented in Chapter 7 along with an overview of different modelling approaches and associated simulation study. We finish with a discussion of future extensions for the application to data.

Part I

Temporal Analysis of Gene Transcription

CHAPTER 2

STOCHASTIC REACTION NETWORKS

We demand rigidly defined areas of doubt and uncertainty.

Douglas Adams, Hitchhiker's Guide to the Galaxy

2.1 Introduction

Stochastic reaction networks (SRNs) can be used to model systems of reactions by Markov jump processes (MJPs). Consider a system of ν stochastic reactions involving D molecular species, $\mathbf{X} = (X_1, ..., X_D)^T$ in a well-mixed environment of volume Ω. The stochastic process can be represented by the set of reactions,

$$\mathcal{P}^{\mathrm{T}}\mathbf{X} \xrightarrow{\mathbf{h}} \mathcal{Q}^{\mathrm{T}}\mathbf{X},$$

for matrices $\mathcal{P}$ and $\mathcal{Q}$. The vector $\mathbf{h}$, is the vector of hazard functions describing the rate at which each reaction occurs and $S := \mathcal{Q}^{\mathrm{T}} - \mathcal{P}^{\mathrm{T}} := [v_1, ..., v_\nu]$ is the stoichiometric matrix. The vectors v_j, describe the corresponding change in state for each reaction j. In general, each hazard function will depend on the state of the system, $\mathbf{x}$, and the associated kinetic rate of the reaction, denoted by θ. By the law

of mass action, the hazard functions are given by,

$$h_j(\mathbf{X}, \theta_j) = \theta_j \prod_{k=1}^{D} \binom{X_k}{\mathcal{P}_{jk}}, \quad \text{for} \quad j = 1, ..., \nu, \tag{2.1}$$

where $\mathcal{P}_{jk}$ is the jkth element of $\mathcal{P}$ and X_k is the kth element of the state vector $\mathbf{X}$.

One can also define the stochastic reaction network in terms of the concentration, $\mathbf{X}^{(\Omega)} := \mathbf{X}/\Omega$, through a classical rescaling (Wilkinson, 2011). In particular, the rescaled hazard rates are defined by,

$$\begin{aligned} h_j^{(\Omega)}\left(\mathbf{X}^{(\Omega)}, \theta_j\right) &= \Omega h_j\left(\mathbf{X}^{(\Omega)}, \theta_j\right) \\ &= \Omega^{o_j - 1} h_j\left(\mathbf{X}, \theta_j\right) \\ &= \theta_j \Omega^{o_j - 1} \prod_{k=1}^{D} \binom{X_k}{\mathcal{P}_{jk}}, \end{aligned} \tag{2.2}$$

where the order of each reaction is defined to be the number of reactant species, $o_j := \sum_{i=1}^{D} \mathcal{P}_{ij}\, \mathbb{I}_{[\mathcal{P}_{ij} \neq 0]}$. Consider the example of a zero order reaction of the form,

$$\emptyset \xrightarrow{\theta} X.$$

The hazard function will be given by θ, and a concentration hazard of $\Omega\theta$. Similarly, a first order reaction of the form,

$$X \xrightarrow{\theta} Y,$$

will have hazard function θX, equivalently, concentration hazard of $\theta X^{(\Omega)} \equiv \theta X/\Omega$.

Example. Consequently, the gene expression model introduced in Chapter 1 can be formulated in the following way,

$$\emptyset \xrightarrow{\beta(t)} M, \qquad M \xrightarrow{\delta_m} \emptyset, \tag{2.3}$$

$$M \xrightarrow{\alpha} M + P, \quad P \xrightarrow{\delta_p} \emptyset. \tag{2.4}$$

Letting $\mathbf{X} = (M, P)^T$ be the vector of states, the associated stoichiometric matrix and hazard rates are given by,

$$S = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}^T, \quad \mathbf{h}(\mathbf{X}, \theta) = \begin{pmatrix} \beta(t) & \delta_m M & \alpha M & \delta_p P \end{pmatrix}^T.$$

The stoichiometric matrix and hazard vector completely specify the reaction network. Specifically, given the system is currently in state $\mathbf{x}$, the probability of reaction j occurring so that the state vector becomes $\mathbf{x}+\mathbf{v}_j$ in the next infinitesimal dt time, is given by $h_j(\mathbf{x},\theta_j)dt$. From this, it is straightforward to derive (Wilkinson, 2011) that the next reaction to occur will be at time $t+\tau$ and of type j with probability,

$$\mathbb{P}(\mathbf{X}(t+\tau)=\mathbf{x}+\mathbf{v}_j|\mathbf{X}(t)=\mathbf{x})=e^{-h_0(\mathbf{x},\theta)\tau}h_j(\mathbf{x},\theta_j), \tag{2.5}$$

where $h_0(\mathbf{x},\theta)=\sum_{j=1}^{\nu}h_j(\mathbf{x},\theta_j)$. This identity forms the basis of the stochastic simulation algorithm (SSA), (see for example, Gillespie (1977)) from which we can generate exact sample paths of a given system.

Given the above properties, stochastic reaction networks can be formulated as a Markov jump process (MJP) (Stathopoulos and Girolami, 2013), where the D dimensional stochastic process $\mathbf{X}(t)=(X_1(t),\ldots,X_D(t))$ satisfies the Markov property that the probability of the current state given its entire history depends only on the state at the previous time point, i.e.

$$\mathbb{P}(\mathbf{X}(t_i)|\mathbf{X}(t_1),\ldots,\mathbf{X}(t_{i-1}))=\mathbb{P}(\mathbf{X}(t_i)|\mathbf{X}(t_{i-1})), \tag{2.6}$$

for any sequence $t_1<\cdots<t_i$ times.

The transition probability is defined for all times s and t, such that $s\leq t$, by $\mathbb{P}(\mathbf{X}(t)=\mathbf{x}_t|\mathbf{X}(s)=\mathbf{x}_s)$ and using the above Markov property, can be shown to satisfy the Chapman-Kolmogorov equations,

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}(t+\tau)=\mathbf{x}|\mathbf{X}(0)=\mathbf{x}_0)\\
&\qquad=\sum_k \mathbb{P}(\mathbf{X}(t+\tau)=\mathbf{x},\,\mathbf{X}(\tau)=k|\mathbf{X}(0)=\mathbf{x}_0)\\
&\qquad=\sum_k \mathbb{P}(\mathbf{X}(t+\tau)=\mathbf{x}|\mathbf{X}(\tau)=k,\mathbf{X}(0)=\mathbf{x}_0)\,\mathbb{P}(\mathbf{X}(\tau)=k|\mathbf{X}(0)=\mathbf{x}_0)\\
&\qquad=\sum_k \mathbb{P}(\mathbf{X}(t+\tau)=\mathbf{x}|\mathbf{X}(\tau)=k)\,\mathbb{P}(\mathbf{X}(\tau)=k|\mathbf{X}(0)=\mathbf{x}_0)\\
&\qquad=\sum_k \mathbb{P}(\mathbf{X}(t)=\mathbf{x}|\mathbf{X}(0)=k)\,\mathbb{P}(\mathbf{X}(\tau)=k|\mathbf{X}(0)=\mathbf{x}_0).
\end{aligned} \tag{2.7}$$

Letting $\mathcal{G}_{kx}(\tau) := \frac{1}{\tau}(\mathbb{P}(\mathbf{X}(\tau) = k|\mathbf{X}(0) = \mathbf{x}_0) - \mathbb{P}(\mathbf{X}(0) = k|\mathbf{X}(0) = \mathbf{x}_0))$, equation (2.7) can be rewritten to give,

$$\begin{aligned}\mathbb{P}(\mathbf{X}(t+\tau) = \mathbf{x}|\mathbf{X}(0) = \mathbf{x}_0) = \mathbb{P}(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = \mathbf{x}_0) \\ +\tau \sum_k \mathbb{P}(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = k)\mathcal{G}_{kx}(\tau),\end{aligned}$$

which can be rearranged to yield,

$$\begin{aligned}\mathbb{P}(\mathbf{X}(t+\tau) = \mathbf{x}|\mathbf{X}(0) = \mathbf{x}_0) - \mathbb{P}(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = \mathbf{x}_0) \\ = \tau \sum_k \mathbb{P}(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = k)\mathcal{G}_{kx}(\tau).\end{aligned}$$

Letting $\tau \downarrow 0$, the generator of the process is given by $\mathcal{G} = (\mathcal{G}_{kx})$, where $\mathcal{G}_{kx} := \lim_{\tau \downarrow 0} \mathcal{G}_{kx}(\tau)$. Consequently, as $\tau \downarrow 0$, we obtain Kolmogorov's forward equations,

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = \mathbf{x}_0) = \sum_k \mathbb{P}(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = k)\mathcal{G}_{kx}. \tag{2.8}$$

For ease of notation, we let $\mathbb{P}(\mathbf{x}, t) := \mathbb{P}(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = \mathbf{x}_0)$ be the transition density from time 0 to time t, subject to the initial condition $\mathbb{P}(\mathbf{x}, 0) = \mathbb{I}_{[\mathbf{x}=\mathbf{x}_0]}$. Kolmogorov's forward equations can then be rewritten (see for example Wilkinson (2011)) in the following form,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}(\mathbf{x}, t) = \sum_{j=1}^{J} h_j(\mathbf{x} - \mathbf{v}_j, \theta_j)\,\mathbb{P}(\mathbf{x} - \mathbf{v}_j, t) - h_j(\mathbf{x}, \theta_j)\,\mathbb{P}(\mathbf{x}, t), \tag{2.9}$$

$$\mathbb{P}(\mathbf{x}, 0) = \begin{cases} 1 \text{ if } \mathbf{x} = \mathbf{x}_0 \\ 0 \text{ if } \mathbf{x} \neq \mathbf{x}_0. \end{cases} \tag{2.10}$$

Equation (2.9) is called the (chemical) master equation (ME) and is often intractable for even very simple systems. Nevertheless, it is useful for deriving approximations to the system and can be used to assess the accuracy of different approximations in the literature (Ferm et al., 2008).

Example. The two species gene transcription model described by equations (2.3) - (2.4) satisfies the following master equation,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}(m, p, t) = \beta(t)\,\mathbb{P}(m-1, p, t) + \delta_m(m+1)\,\mathbb{P}(m+1, p, t) \\ + \alpha m\,\mathbb{P}(m, p-1, t) + \delta_p(p+1)\,\mathbb{P}(m, p+1, t) \\ - (\beta(t) + \delta_m m + \alpha m + \delta_p p)\,\mathbb{P}(m, p, t).\end{aligned} \tag{2.11}$$

$$\mathbb{P}(m,p,0) = \begin{cases} 1 & \text{if} \quad m = m_0 \quad \text{and} \quad p = p_0, \\ 0 & \text{if} \quad m \neq m_0 \quad \text{or} \quad p \neq p_0. \end{cases} \tag{2.12}$$

Here, $\mathbb{P}(m,p,t)$ is the transition probability that at time t, the number of mRNA molecules (M) is given by m and the number of proteins (P) is given by p, subject to the initial condition of m_0 mRNA molecules and p_0 proteins at time $t = 0$.

2.2 Exact Inference

If complete data on all species and reactions were available, inference would be straightforward since the likelihood is then given by,

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} h_{j_i}(\mathbf{x}(t_i), \theta_j) \ \exp\bigg(- \sum_{i=1}^{n} h_0(\mathbf{x}(t_i), \theta)[t_i - t_{i-1}] \bigg), \tag{2.13}$$

where n is the number of reactions that take place, $j_1, ..., j_n$ is the sequence of reaction types and $t_1, .., t_n$ are the associated timings of each reaction. However, in molecular biology complete data paths are rarely available and commonly only a subset of species are measured with error.

There are therefore two broad approaches one can take to perform inference. Either one can integrate over the latent reaction paths between observations or one can work with (approximate) transition densities of the system. We focus our methods on the latter approach, which we will argue is computationally more feasible although has the disadvantage that a likelihood approximation is often necessary to accomplish this. The former approach allows one to work with the exact system and attention has been focused on performing these high-dimensional integrations in a computationally efficient way and is reviewed below.

Andrieu et al. (2010) show how particle MCMC methods can be used to perform inference on MJPs, in particular the stochastic kinetic Lotka-Volterra model although it was found to perform poorly in low measurement error scenarios (Golightly and Wilkinson, 2011). Other approaches for inference on the exact system include a simulation based method (Amrein and Künsch, 2012), a reversible jump MCMC method (Boys et al., 2008), an implementation of uniformisation (Choi and Rempala, 2012) and the MCEM^2 of Daigle et al. (2012), which uses techniques for simulating rare events. Two recent examples that apply their methods to real data are the delayed acceptance MCMC method of Golightly et al. (2014) applied to epidemic data and

the dynamic prior propagation method of Zechner et al. (2014) who model an artificially controlled gene expression system in yeast. The delayed acceptance method of Golightly et al. (2014) is again an application of particle MCMC methods. However, in order to decrease the computational cost, sample paths are first proposed under a fast approximation and only if these are accepted is a sequential Monte Carlo scheme used to estimate the true latent states. In this way, their algorithm avoids computing proposals under the true likelihood that are likely to be rejected.

The dynamic prior propagation method of Zechner et al. (2014) is based on a hierarchical structure that allows one to construct a marginal reaction network over the rate constants that vary between cells. In this way, the method only infers sufficient statistics of the distribution of the rate constants rather than each individual cell specific constant. The authors successfully applied their approach to single cell time course measurements subject to white noise error. Although described as a scalable approach, no indication of computational cost is given. Moreover, it is clear that although the method scales well as the number of individual cells increases, it is less clear how well the methods will scale as the number of observations per cell increases, particularly since it relies on an SMC update through stochastic simulation of the latent states.

One further approach to inference on the exact MJP has been achieved through approximate inference techniques. In these scenarios one continues to work with the exact stochastic system but obtains only an approximation to the true posterior density. Examples include ABC (Approximate Bayesian Computation, Beaumont et al. (2002)) and Variational Bayes. ABC methods have been applied to the stochastic Lotka-Volterra model (Fearnhead and Prangle, 2012), a stochastic autologistic model (Drovandi and Pettitt, 2011) and gene network models (Lillacci and Khammash, 2013). In contrast, Variational Bayes techniques have been used to infer the logics of two competing transcription factor interactions within a reduced MJP via the optimisation of the Kullback-Leibler divergence between the true process and a known family of densities (Opper and Sanguinetti, 2010).

Each of the methods mentioned above have their own disadvantages. Although approximate inference techniques can be useful for inferring parameters, it is not always clear how the approximate posterior is related to the true posterior. Recently, Owen et al. provided a comparison of ABC and exact inference techniques applied to stochastic reaction networks. They found the exact approaches to be preferable when observed times series were relatively long and moreover, found ABC performed poorly in the presence of measurement error. Having said this, the exact inference

methods also have significant limitations in a practical framework. Without exception, all the exact inference techniques reviewed here assume a known scaling factor of 1 and often also known measurement error. Moreover, the techniques used are generally computationally burdensome with respect to the size of data we consider (a typical dataset in our experimental framework consists of 100 time series of approximately 190 measurements per time series). In a molecular biology framework experimental methods will invariably result in a measurement process with both unknown error and scaling as the direct number of molecules is not observable. One approach is to rescale the data based on additional experiments (Zechner et al., 2014) as the incorporation of both unknown measurement error and scaling is non-trivial in this framework and which we will consider in some detail.

Further details of the exact approaches to inference mentioned here are reviewed in the supplementary material of Appendix A.

2.3 Likelihood Approximations

In contrast to the above approaches, we consider the feasibility of approximating the underlying MJP. This is achieved by approximating the transition densities, $\mathbb{P}(\mathbf{x}, t)$, that solve the, rarely tractable, chemical master equation given in (2.9).

In order to define the limiting approximations, we let $\mathbf{X}^{(\Omega)} := \mathbf{X}/\Omega$ denote the concentration of $\mathbf{X}$, and in particular, $\mathbf{X}^{(\Omega)}$ satisfies the same SRN as $\mathbf{X}$ with hazard rates given by $\mathbf{h}^{(\Omega)}$ as defined in (2.2).

It is useful in the following to express the Markov jump process in terms of the following Poisson process. Namely, for fixed τ the following equality holds,

$$\begin{aligned} \mathbf{X}(t+\tau) &= \mathbf{X}(t) + \sum_{j=1}^{\nu} \mathbf{v}_j \mathcal{K}_j, \\ \mathcal{K}_j &\sim Pois\bigg(\int_t^{t+\tau} h_j(\mathbf{X}(s), \theta_j) \ \mathrm{d}s \bigg), \end{aligned} \tag{2.14}$$

where $\mathcal{K}_j$ is the number of type j events occurring in the interval. Equivalently, the rescaled concentration process $\mathbf{X}^{(\Omega)}$ satisfies the Poisson process,

$$\begin{aligned} \mathbf{X}^{(\Omega)}(t+\tau) &= \mathbf{X}^{(\Omega)}(t) + \tfrac{1}{\Omega} S \mathcal{K}^{(\Omega)}, \\ \mathcal{K}^{(\Omega)} &\sim Pois\bigg(\int_t^{t+\tau} \Omega \mathbf{h}(\mathbf{X}^{(\Omega)}(s), \theta) \ \mathrm{d}s \bigg), \end{aligned} \tag{2.15}$$

where S is the stoichiometric matrix, $\mathcal{K}^{(\Omega)}$ is a vector of Poisson random variables and $\mathbf{h} = (h_1, ..., h_\nu)^T$ is the vector of hazard rates.

2.3.1 Macroscopic Limit

The direct deterministic analogue, $\mathbf{X}^{\mathrm{D}}(t)$, of the stochastic model, $\mathbf{X}(t)$, is given by the conditional expectation of the stochastic process given its history (Chesson, 1978),

$$\mathbf{X}^{\mathrm{D}}(t+\tau) = \mathbb{E}\left[\mathbf{X}(t+\tau)|\mathbf{X}(t) = \mathbf{X}^{\mathrm{D}}(t)\right] = \mathbf{X}^{\mathrm{D}}(t) + \int_t^{t+\tau} S\mathbf{h}(\mathbf{X}^{\mathrm{D}}(s), \theta)\ \mathrm{d}s. \quad (2.16)$$

This can be rewritten in the equivalent ODE for $\mathbf{X}^{\mathrm{D}}$, often referred to as the reaction rate equation (RRE) or macroscopic limit,

$$\frac{\mathrm{d}\mathbf{X}^{\mathrm{D}}}{\mathrm{d}t} = A(\mathbf{X}^{\mathrm{D}}) := \sum_{j=1}^{\nu} \mathbf{v}_j h_j(\mathbf{X}^{\mathrm{D}}, \theta_j) = S\mathbf{h}(\mathbf{X}^{\mathrm{D}}, \theta), \quad (2.17)$$

$$\mathbf{X}^{\mathrm{D}}(0) = \mathbf{x}_0. \quad (2.18)$$

Kurtz (1970) and Anderson and Kurtz (2011) show, using the law of large numbers, that this ODE can be derived as the limit of $\mathbf{X}^{(\Omega)}$ as $\Omega \to \infty$,

$$\mathbf{X}^{(\Omega)}(t+\tau) \to \mathbf{X}^{(\Omega)}(t) + \int_t^{t+\tau} S\mathbf{h}(\mathbf{X}^{(\Omega)}(s), \theta)\ \mathrm{d}s =: \mathbf{X}^{\mathrm{D}}(t+\tau), \quad \text{as } \Omega \to \infty.$$

As a consequence, for sufficiently large systems, intrinsic variability can be assumed to vanish and a deterministic ODE model can be used. This approach may be appropriate for modelling high-level aggregate data with negligible intrinsic variability as in Jenkins et al. (2013).

For example, the reaction rate equations of the gene transcription model (2.3)-(2.4) are given by,

$$M^{\mathrm{D}}(t+\tau) = M^{\mathrm{D}}(t) + \int_t^{t+\tau} \beta(s) - \delta_m M^{\mathrm{D}}(s)\ \mathrm{d}s,$$
$$P^{\mathrm{D}}(t+\tau) = P^{\mathrm{D}}(t) + \int_t^{t+\tau} \alpha M^{\mathrm{D}}(s) - \delta_p P^{\mathrm{D}}(s)\ \mathrm{d}s.$$

Equivalently, formulating as an ODE,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} M^{\mathrm{D}}(t) \\ P^{\mathrm{D}}(t) \end{pmatrix} = \begin{pmatrix} \beta(t) - \delta_m M^{\mathrm{D}}(t) \\ \alpha M^{\mathrm{D}}(t) - \delta_p P^{\mathrm{D}}(t) \end{pmatrix}.$$

2.3.2 Limiting Diffusion Processes

In this section, we consider two stochastic approximations to the MJP, the chemical Langevin equation (CLE) and the linear noise approximation (LNA). Both give rise to systems of SDEs but the literature often refers to the CLE as the diffusion approximation or DA, we choose to avoid this terminology to avoid confusion.

Chemical Langevin Equation

The chemical Langevin equation was first derived in the chemical physics literature in Gillespie (2000). We will follow this heuristic derivation although a more rigorous treatment can be found in Anderson and Kurtz (2011).

Assuming τ is chosen to be small enough such that $h_j(\mathbf{X}^{(\Omega)}, \theta_j)$ can be considered constant over the interval $[t, t+\tau)$ $\forall j$, known as the *first leap condition* (Gillespie, 2000), the updating equation (2.15) becomes,

$$\begin{aligned} &\mathbf{X}^{(\Omega)}(t+\tau) \approx \mathbf{X}^{(\Omega)}(t) + \tfrac{1}{\Omega} S \tilde{\mathcal{K}}^{(\Omega)}, \\ &\tilde{\mathcal{K}}^{(\Omega)} \sim Pois(\Omega \mathbf{h}(\mathbf{X}^{(\Omega)}(t), \theta)\tau), \end{aligned} \tag{2.19}$$

where the integrand in the Poisson rate has been replaced by a constant. Under the *second leap condition*, namely $\Omega h_j(\mathbf{X}^{(\Omega)}(t), \theta_j)\tau$ is large $\forall j$, this Poisson random variate can be replaced by a normal density yielding, the (chemical) Langevin equation,

$$\begin{aligned} &\mathbf{X}^{\mathrm{C}}(t+\tau) = \mathbf{X}^{\mathrm{C}}(t) + \tau S \mathbf{h}(\mathbf{X}^{\mathrm{C}}, \theta) + \sqrt{\tau}\sqrt{S \operatorname{diag}\left(\mathbf{h}(\mathbf{X}^{\mathrm{C}}, \theta)\right) S^T}\, \mathcal{Z}, \\ &\mathcal{Z} \sim N(0, I_D). \end{aligned} \tag{2.20}$$

Formulating the CLE in terms of the concentration process $\mathbf{X}^{(\Omega)}$, we notice that the second leap condition is satisfied as the system size $\Omega \to \infty$.

Taking the limit as $\tau \to 0$, equation (2.20) can be expressed by the following Itô

diffusion process (Gardiner, 1985),

$$\mathrm{d}\mathbf{X}^{\mathrm{C}} = A\left(\mathbf{X}^{\mathrm{C}}\right)\ \mathrm{d}t + B\left(\mathbf{X}^{\mathrm{C}}\right)\ \mathrm{d}\mathbf{W}_t, \tag{2.21}$$

$$A\left(\mathbf{X}^{\mathrm{C}}\right) := \sum_{j=1}^{\nu} \mathbf{v}_j h_j\left(\mathbf{X}^{\mathrm{C}}, \theta_j\right) = S\mathbf{h}\left(\mathbf{X}^{\mathrm{C}}, \theta\right), \tag{2.22}$$

$$B\left(\mathbf{X}^{\mathrm{C}}\right) := \sqrt{S\,\mathrm{diag}\left(\mathbf{h}\left(\mathbf{X}^{\mathrm{C}}, \theta\right)\right) S^T}. \tag{2.23}$$

Consequently, in the limit as $\tau \to 0$, the CLE will preserve the first and second moments of the true Poisson process in (2.14) albeit through a Gaussian approximation.

Returning to the linear gene transcription model of (2.3)-(2.4) and letting $\mathbf{X}^{\mathrm{C}} := (M^{\mathrm{C}}, P^{\mathrm{C}})^T$, the corresponding CLE will satisfy (2.21) with,

$$A\left(\mathbf{X}^{\mathrm{C}}\right) = \begin{pmatrix} \beta(t) - \delta_m M^{\mathrm{C}}(t) \\ \alpha M^{\mathrm{C}}(t) - \delta_p P^{\mathrm{C}}(t) \end{pmatrix},$$

$$B\left(\mathbf{X}^{\mathrm{C}}\right) = \begin{pmatrix} \sqrt{\beta(t) + \delta_m M^{\mathrm{C}}(t)} & 0 \\ 0 & \sqrt{\alpha M^{\mathrm{C}}(t) + \delta_p P^{\mathrm{C}}(t)} \end{pmatrix}.$$

The CLE has been used extensively for inference within SRNs, (Golightly and Wilkinson, 2005, 2011; Heron et al., 2007). Despite this there are several drawbacks, not least of all, that the transition density often remains intractable and one often works with a discretisation, for example the Euler-Maruyama approximation. Moreover, in practice, data are measured at discrete time intervals that cannot be assumed to satisfy the first leap condition and consequently, one is often required to bridge between observations. Various approaches have been used to impute latent points between observations to ensure that this condition is satisfied. Heron et al. (2007) used a bridging approach (Elerian et al., 2001), whereas Golightly and Wilkinson (2011) employ a particle MCMC scheme to impute these points, however in both cases, the parameters of the measurement equation were assumed to be known.

Linear Noise Approximation

The linear noise approximation (LNA) is a linearisation of the master equation and always results in analytical transition densities. Derivations of varying degrees of rigour can be found in numerous sources (see for example, van Kampen (1961);

Kurtz (1971); Wallace et al. (2012)). Van Kampen's system size expansion evolves from the ansatz,

$$\mathbf{X}^{\mathrm{L}}(t) = \phi(t) + \Omega^{-1/2}\xi(t), \tag{2.24}$$

where ϕ is a deterministic path, ξ a stochastic fluctuation and Ω is the size of the system. In particular, $\phi := \mathbf{X}^{\mathrm{D}}$, is the macroscopic solution. The derivation then proceeds via a second order Taylor expansion about the master equation for the solution (2.24). Kurtz (1971), on the other hand, provides a rigorous foundation for the LNA with a detailed application to SRNs given in Anderson and Kurtz (2011), which we follow here. In particular, the LNA is derived as a central limit theorem to the Poisson process given in equation (2.15). To see this, we consider $V^{(\Omega)} := \sqrt{\Omega}(\mathbf{X}^{(\Omega)} - \phi)$ to be the deviation between the Poisson process, $\mathbf{X}^{(\Omega)}$, and the macroscopic limiting process, ϕ,

$$\begin{aligned}
&V^{(\Omega)}(t+\tau) \\
&\quad = \sqrt{\Omega}(\mathbf{X}^{(\Omega)}(t+\tau) - \phi(t+\tau)) \\
&\quad = \sqrt{\Omega}(\mathbf{X}^{(\Omega)}(t) - \phi(t)) + \sqrt{\Omega}\left(\tfrac{1}{\Omega} S\mathcal{K}^{(\Omega)} - S\int_t^{t+\tau} \mathbf{h}(\phi(s),\theta)\ \mathrm{d}s\right) \\
&\quad = V^{(\Omega)}(t) + \tfrac{1}{\sqrt{\Omega}} S\mathcal{K}^{(\Omega)} - \sqrt{\Omega}\int_t^{t+\tau} A(\phi(s))\ \mathrm{d}s,
\end{aligned}$$

where A is defined as in equation (2.17). Thus we have,

$$\begin{aligned}
V^{(\Omega)}(t+\tau) - V^{(\Omega)}(t) &= \tfrac{1}{\sqrt{\Omega}} S\mathcal{K}^{(\Omega)} - \sqrt{\Omega}\int_t^{t+\tau} A(\phi(s))\ \mathrm{d}s, \\
\mathcal{K}^{(\Omega)} &\sim Pois\left(\int_t^{t+\tau} \Omega\mathbf{h}(\mathbf{X}^{(\Omega)}(s),\theta)\ \mathrm{d}s\right).
\end{aligned} \tag{2.25}$$

Using the trick of adding zero, we have,

$$\begin{aligned}
V^{(\Omega)}(t+\tau) - V^{(\Omega)}(t) &= \tfrac{1}{\sqrt{\Omega}} S\mathcal{K}^{(\Omega)} - \sqrt{\Omega}\int_t^{t+\tau} S\mathbf{h}(\mathbf{X}^{(\Omega)}(s),\theta)\ \mathrm{d}s \\
&\quad + \sqrt{\Omega}\int_t^{t+\tau} S\mathbf{h}(\mathbf{X}^{(\Omega)}(s),\theta)\ \mathrm{d}s - \sqrt{\Omega}\int_t^{t+\tau} A(\phi(s))\ \mathrm{d}s, \\
&= \tfrac{1}{\sqrt{\Omega}} S\mathcal{K}^{(\Omega)} - \sqrt{\Omega}\int_t^{t+\tau} S\mathbf{h}(\mathbf{X}^{(\Omega)}(s),\theta)\ \mathrm{d}s \\
&\quad + \sqrt{\Omega}\int_t^{t+\tau} \left(A(\mathbf{X}^{(\Omega)}(s)) - A(\phi(s))\right)\ \mathrm{d}s.
\end{aligned}$$

By the central limit theorem, $\frac{1}{\sqrt{\Omega}} S\mathcal{K}^{(\Omega)} - \sqrt{\Omega}\int_t^{t+\tau} S\mathbf{h}(\mathbf{X}^{(\Omega)}(s),\theta)\ \mathrm{d}s \to \mathcal{Z}^{(\Omega)}$, as $\Omega \to \infty$, where,

$$\mathcal{Z}^{(\Omega)} \sim N\bigg(0,\ S\int_t^{t+\tau} \mathbf{h}(\mathbf{X}^{(\Omega)}(s),\theta)\ \mathrm{d}s\, S^T\bigg). \tag{2.26}$$

Moreover, under the assumptions that there exists a unique solution to the initial value problem (2.17) and that the hazard rates are multinomial (assumptions that are immediately satisfied by a stochastic reaction network), theorem K of Barbour (1974) allows one to rewrite the following,

$$\begin{aligned}\sqrt{\Omega}\int_t^{t+\tau} \big(A(\mathbf{X}^{(\Omega)}(s)) - A(\phi(s))\big)\ \mathrm{d}s &\approx \sqrt{\Omega}\int_t^{t+\tau} \big(\nabla A(\phi(s))(\mathbf{X}^{(\Omega)}(s) - \phi(s))\big)\ \mathrm{d}s \\ &= \int_t^{t+\tau} \big(J(\phi(s))V^{(\Omega)(s)}\big)\ \mathrm{d}s,\end{aligned}$$

where $J(\phi(s))$ is the Jacobian, $J_{ij} := \frac{\partial A_j}{\partial \phi_i}$, of the macroscopic ODE. Consequently, taking the limit as $\Omega \to \infty$, and defining $\xi(t) := \lim_{\Omega\to\infty} V^{(\Omega)}(t) = \lim_{\Omega\to\infty}(\sqrt{\Omega}(\mathbf{X}^{(\Omega)}(t) - \phi(t)))$, equation (2.25) becomes,

$$\begin{aligned}\xi(t+\tau) - \xi(t) &= \int_t^{t+\tau} J(\phi(s))\xi(s)\ \mathrm{d}s + \mathcal{Z}, \\ Z &\sim N\bigg(0,\ \int_t^{t+\tau} B(\phi(s))B(\phi(s))^T\ \mathrm{d}s\bigg),\end{aligned}$$

where B is as in (2.23) and comes from equation (2.26).

We therefore arrive at the full specification of the LNA,

$$\mathbf{X}^{\mathrm{L}}(t) = \phi(t) + \Omega^{-1/2}\xi(t), \tag{2.27}$$

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = A(\phi(t)), \tag{2.28}$$

$$\mathrm{d}\xi = J(\phi(t))\xi(t)\ \mathrm{d}t + B(\phi(t))dW_t, \tag{2.29}$$

where dW_t are independent Wiener processes. Equation (2.29) is linear with Itô representation and thus the transition, $\mathbb{P}(\xi(t+\tau)|\xi(t))$, is Gaussian with mean and variance defined by (Komorowski et al., 2009),

$$\begin{aligned}\frac{\mathrm{d}\mu}{\mathrm{d}t} &= J(\phi(t))\mu(t), \\ \frac{\mathrm{d}\Sigma}{\mathrm{d}t} &= \Sigma(t)J(\phi(t))^T + J(t)\Sigma(t)^T + B(\phi(t))B(\phi(t))^T.\end{aligned}$$

Correspondingly, the transition probabilities of the state vector are derived to be (Finkenstädt et al., 2013),

$$\mathbb{P}\left(\mathbf{X}^{\mathrm{L}}(t+\tau)|\mathbf{X}^{\mathrm{L}}(t)\right) = N\left(\phi(t) + \Omega^{-1/2}\mu(t+\tau), \Omega^{-1}\Sigma(t+\tau)\right).$$

Given this formulation, the LNA can always be expressed as a linear Gaussian state space model. For example, let $\mathbf{X}^{\mathrm{L}}(t)$ be the LNA of a linear stochastic reaction system. In this case, the Jacobian is independent of ϕ and consequently is constant in time so that the state space representation takes the form,

$$\begin{aligned}
\mathbf{X}^{\mathrm{L}}(t+\tau) &= e^{J\tau}\mathbf{X}^{\mathrm{L}}(t) + \left(\phi(t+\tau) - e^{J\tau}\phi(t)\right) + \Omega^{-1/2}\eta(t+\tau), \qquad (2.30)\\
\eta(t+\tau) &\sim N(0, \Sigma(t+\tau))\\
\Sigma(t+\tau) &= \int_t^{t+\tau} \left[e^{J(t+\tau-s)}B(s)\right]\left[e^{J(t+\tau-s)}B(s)\right]^T \,\mathrm{d}s.
\end{aligned}$$

As before, ϕ denotes the solution to the macroscopic ODE and J is the associated Jacobian. To be explicit, for the gene transcription model (2.3)-(2.4), $\phi := (\phi_m, \phi_p)^T$, where ϕ_m and ϕ_p solve the following ODE system,

$$\begin{aligned}
\frac{\mathrm{d}\phi_m}{\mathrm{d}t} &= \beta(t) - \delta_m\phi_m(t),\\
\frac{\mathrm{d}\phi_p}{\mathrm{d}t} &= \alpha\phi_m(t) - \delta_p\phi_p(t),
\end{aligned}$$

and

$$J = \begin{pmatrix} -\delta_m & 0 \\ \alpha & -\delta_p \end{pmatrix}, \qquad B(\phi(t)) = \begin{pmatrix} \sqrt{\beta(t) + \delta_m\phi_m(t)} & 0 \\ 0 & \sqrt{\alpha\phi_m(t) + \delta_p\phi_p(t)} \end{pmatrix}.$$

Both the CLE and LNA are derived in the limit as the system size $\Omega \to \infty$ with precise statements given in Kurtz (1971, 1978). Despite the LNA commonly being derived as an approximation to the CLE, Anderson and Kurtz (2011) show that in fact less stringent assumptions are required for the derivation and argue that the foundations of the LNA are well justified whereas the CLE is derived less rigorously. Inference on different transcription networks including autoregulatory and dimerisation systems using the LNA are given in Ruttor and Opper (2009); Komorowski et al. (2009); Stathopoulos and Girolami (2013); Finkenstädt et al. (2013) and Fearnhead et al. (2014). Although the LNA is derived in the large system size limit, these studies found reasonable performance for mesoscopic systems.

2.3.3 Birth Death Approximation

In addition to the already existing approximations given above, one can construct an approximate reaction network specifically for the transcription model (2.3)-(2.4). In particular, we consider an approximation consisting of conditionally independent birth-death networks,

$$\emptyset \xrightarrow{\beta(t)} M, \quad M \xrightarrow{\delta_m} \emptyset, \tag{2.31}$$

$$\emptyset \xrightarrow{\alpha M^*(t)} P, \quad P \xrightarrow{\delta_p} \emptyset, \tag{2.32}$$

which corresponds to the following factorisation of the joint transition density,

$$\begin{aligned}\mathbb{P}(M(t), P(t)|M(0), P(0)) &= \mathbb{P}(M(t)|M(0), P(0))\, \mathbb{P}(P(t)|M(0), M(t), P(0)) \\ &\approx \mathbb{P}(M(t)|M(0))\, \mathbb{P}(P(t)|M^*(t), P(0)).\end{aligned} \tag{2.33}$$

The approximation of this birth-death decomposition arises through the process M^*. Note that the exact system will be derived by setting M^* to be the continuous time mRNA process, $M(t)$. We have considered the three following different definitions of M^*,

M1: $M^*(t) := m_0$, the mRNA level at the previous time point. As the distance between observations becomes small, this will converge to the exact process.

M2: $M^*(t) := m_t$, the mRNA level at the current time point. Again, as the distance between observations becomes small, this will converge to the exact process.

M3: $M^*(t) := \mathbb{E}(M(t)|M(0) = m_0)$, the expected value of the continuous time mRNA process given the previous observation. This approximation will converge to the exact process when either the time between observations becomes small or the intrinsic variability of the mRNA process vanishes.

Note that in all cases, the marginal transition for the mRNA process will remain exact and it is through the protein process that the approximation to the joint transition occurs. Figure 2.1 shows how each approximation M^* compares to the true underlying process M for two different situations. Clearly, approximation M3 underestimates the variance in the true process, with only minor differences being observed between the other two approximations. Figure 2.2 shows the 95% pointwise simulation intervals for the corresponding protein population levels under different

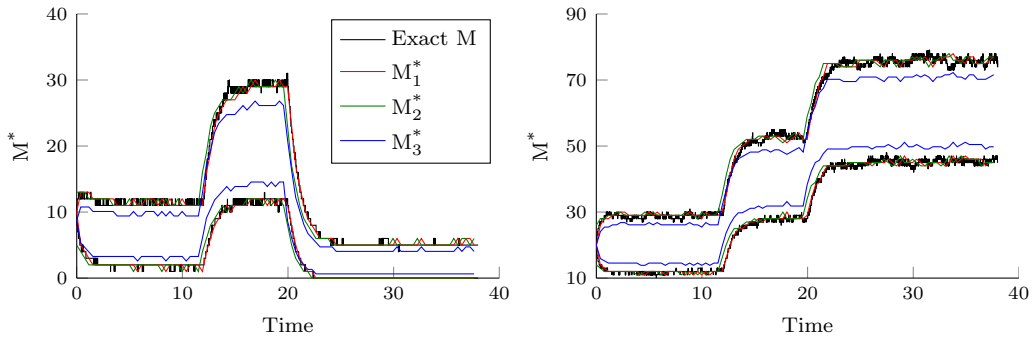


Figure 2.1: In each panel, the black envelope is the empirical 95% pointwise confidence envelope for the true continuous time mRNA process and the red, green and blue envelopes correspond to the empirical 95% pointwise envelopes for M^* under approximation M1, M2, and M3 respectively. The two panels correspond to different simulated scenarios of different sampling intervals and different molecular numbers (increasing from a) to b)). In particular, transcription has been simulated to have switches at 12 and 20 hours in each scenario.

simulations for each of the above approximations. Again, it can be seen that under approximation M3, the variance is underestimated in all scenarios. In comparison, when the sampling interval is small with respect to the speed of the reactions (Figure 2.2 a)), the difference between approximations M1 and M2 is small. However, as the sampling interval increases, some differences appear, particularly about the switch points in transcription. Taking the previous mRNA population as a proxy to the continuous time mRNA process will result in delaying the estimated switch points in transcription, whereas the current mRNA population will accelerate the estimation of switch points. Figure 2.3, which shows the marginal transition density of the protein process under each approximation, shows more clearly that as the sampling interval becomes even larger, approximation M2 becomes the more accurate proxy to the true process. Therefore, we restrict our research to using only approximation M2, which we will term the birth-death decomposition (BDD), and note here that approximation M1 may also yield valid inference for systems with a "reasonable" sampling window.

To solve the system (2.31)-(2.32), we note that a birth-death process of the form,

$$\emptyset \xrightarrow{b(t)} X, \quad X \xrightarrow{d(t)} \emptyset, \tag{2.34}$$

has corresponding master equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}(x,t) = b(t)\,\mathbb{P}(x-1,t) + d(t)\,\mathbb{P}(x+1,t) - (b(t)+d(t))\,\mathbb{P}(x,t), \tag{2.35}$$

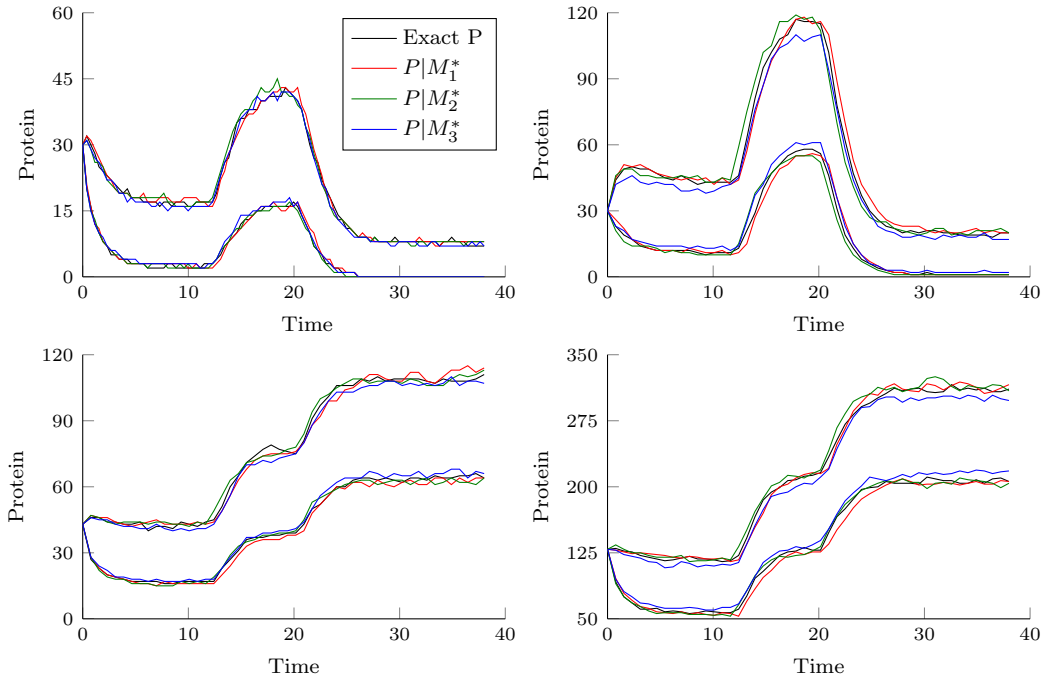


Figure 2.2: In each panel, the black envelope is the empirical 95% pointwise confidence envelope for the true protein process and the red, green and blue envelopes correspond to the empirical 95% pointwise envelopes for the approximate protein process given M^* calculated under approximation M1, M2, and M3 respectively. Each panel corresponds to a different simulated scenario of different sampling intervals and different molecular numbers (increasing from a) to d)). In particular, transcription has been simulated to have switches at 12 and 20 hours in each scenario.

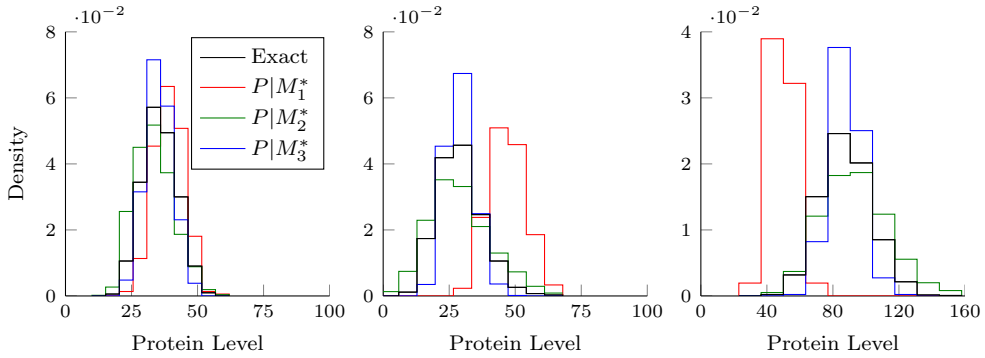


Figure 2.3: In each panel, the black histogram is the empirical transition density for the true marginal protein process and the red, green and blue histograms correspond to the empirical transition densities for the marginal protein process given M^* calculated under approximation M1, M2, and M3 respectively. Each panel corresponds to a different simulated scenario of different sampling intervals (increasing sampling interval from left to right) whilst maintaining the same rate parameter values (same molecular levels). In particular, the final panel has a sampling interval large enough to contain a switch in transcription.

$$\mathbb{P}(x,0) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{if } x \neq x_0, \end{cases}$$

that has a closed form solution (see Appendix B.1 for details). Explicitly, letting Z denote the random variable with transition density satisfying (2.35) then,

$$Z = Z_P + Z_B,$$

where $Z_P \sim Pois(\lambda)$ and $Z_B \sim Bin(x_0, \pi)$ and the corresponding parameters satisfy the following system of ODEs,

$$\begin{aligned} \frac{\mathrm{d}\lambda}{\mathrm{d}t} &= b(t) - d(t)\lambda(t), \quad \lambda(0) = 0, \\ \frac{\mathrm{d}\pi}{\mathrm{d}t} &= -d(t)\pi(t), \qquad \pi(0) = 1. \end{aligned}$$

A full derivation of this result is given in Gardiner (1985). This birth-death decomposition or BDD may be further approximated by replacing the Poisson-binomial convolution by a bivariate normal density truncated to the positive real line so that the transition densities are given by,

$$M(t)|M(0) \sim N_T(\lambda^m + m_0\pi^m, \lambda^m + m_0\pi^m(1-\pi^m)) \tag{2.36}$$

$$P(t)|(M^*(t), P(0)) \sim N_T(\lambda^p + p_0\pi^p, \lambda^p + p_0\pi^p(1-\pi^p)) \tag{2.37}$$

where N_T indicates the normal density truncated to the positive real line and,

$$\frac{\mathrm{d}\lambda^m}{\mathrm{d}t} = \beta(t) - \delta_m\lambda^m(t), \quad \frac{\mathrm{d}\pi^m}{\mathrm{d}t} = -\delta_m\pi^m(t), \quad \lambda^m(0) = 0, \quad \pi^m(0) = 1. \tag{2.38}$$

$$\frac{\mathrm{d}\lambda^p}{\mathrm{d}t} = \alpha m_0 - \delta_p\lambda^p(t), \quad \frac{\mathrm{d}\pi^p}{\mathrm{d}t} = -\delta_p\pi^p(t), \quad \lambda^p(0) = 0, \quad \pi^p(0) = 1. \tag{2.39}$$

This truncated normal approximation to the BDD is then termed the birth-death approximation or BDA.

Although this method may be more widely applied, since many stochastic reaction networks can be expressed as a sequence of conditionally independent birth-death subsystems, we restrict our attention to the gene transcription model that motivates this research.

2.3.4 Other Approximations

So far, we have presented only a selection of approaches that can be found for approximating the exact Markov jump process. Due to the ever increasing interest in modelling biological processes through MJPs, or equivalently stochastic reaction networks, there has been a plethora of approximation techniques developed in the literature within recent years. In general, these approximations can be categorised into two main approaches (Kazeev et al., 2014). Firstly, there are the approximations obtained asymptotically, which include the macroscopic rate equation, the CLE and the LNA as already discussed. Other examples include the Moment Closure approximations (Gomez-Uribe and Verghese, 2007; Ferm et al., 2008; Ullah and Wolkenhauer, 2009), quasi-steady state approaches (Rao and Arkin, 2003) and extensions to the LNA (Grima, 2010; Thomas et al., 2012, 2014).

The second broad approach to approximating the exact MJP is through a truncation of the state space of the system. For example, the finite state projection (FSP) method (Munsky and Khammash, 2006, 2008) allows one to approximate the solution to the master equation to any prespecified degree of accuracy. An alternative truncation approach is through the uniformisation method (Hobolth and Stone, 2009), implemented in the stochastic reaction framework by Choi and Rempala (2012).

Although, in general these alternative approximations within the literature have the advantage over both exact inference methods and the Langevin approximation, in that an analytical expression for the transition density is available, the form of the density still provides many challenges from an inferential viewpoint in the presence of measurement error. Specifically, in the presence of measurement error, the data likelihood remains intractable. To date, there has been only limited research in the area of inference of these additional approximations, specifically only of the truncation methods. Moreover, the inclusion of a measurement process has not, to date, been considered. The issues surrounding the inference problem for these approximations are similar to those facing the BDA, where an analytical transition density is available but not a data likelihood. This is of course a standard problem within state space models and is considered in detail in Chapter 3. We do not consider the inference problem for these alternative approximations but note that future research in this area is likely to be conducive to the field, especially since it is desirable to have both a comparison of the approximation accuracy and a comparison of the amenability and accuracy of the inference problem.

A detailed review of this literature regarding the different approaches to approximating the underlying SRN is given in the supplementary material of Appendix A. We do not consider them any further either for inference or simulation based comparison for several reasons. The Moment Closure and quasi-steady state approaches have both been shown to have limited validity (Schnoerr et al., 2014; Thomas et al., 2012). The slow-scale LNA (Thomas et al., 2012) and conditional LNA (Thomas et al., 2014) are both derived by making explicit assumptions about the regulation of the promoter, however, for our application to the Prolactin gene, little is known about regulation and consequently, these approaches are less amenable. The remaining truncation approaches can be difficult to implement in an efficient way since the truncated state space can remain very large and have limited applicability to long time series.

2.4 Comparative Study

The aim of this section is to explore various features of the different approximations one can use on stochastic reaction networks. For the purpose of simulation, we compare the performance of the exact process to the CLE, the LNA, the BDD and the BDA. We note, that when viewing the inference problem within the following chapter, we will focus attention to the LNA and BDA which have analytical transition densities.

In order to compare these different processes, we apply them to the gene transcription model of (2.3)-(2.4) and investigate both the behaviour of individual transition densities and the accuracy of simulations.

Transition Density

Shown in Figure 2.4 are the corresponding transition densities calculated under the exact process, the CLE, the LNA, the BDD and the BDA. Since an analytical form for the transition density is unavailable for either the exact process or the CLE, these have been calculated empirically via simulation. The most notable differences occur at small molecular numbers where both the CLE and the LNA give non-zero probability to negative populations. Moreover, the BDA and BDD, better capture the asymmetry in the exact density at these low levels.

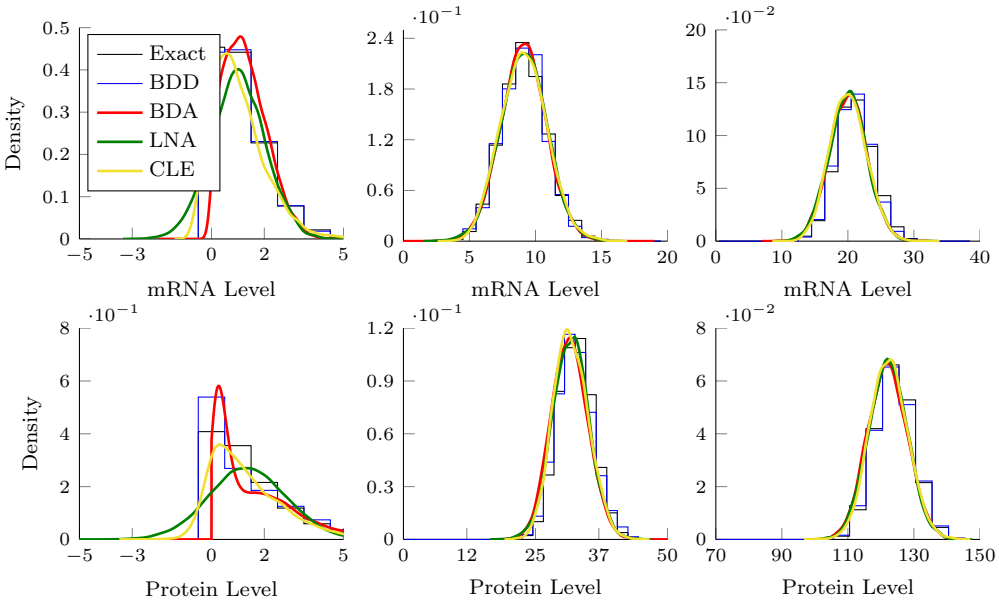


Figure 2.4: Transition densities calculated for the exact process (obtained empirically and shown in grey), the BDD (blue), the truncated normal BDA (red), the LNA (green) and the CLE (yellow). These have been calculated for three different scenarios for both mRNA (top) and for protein (bottom) involving increasing molecular abundances for both species from left to right.

Simulations

The improved precision of the BDA over the LNA becomes apparent in Figure 2.5, which shows the 95% pointwise confidence envelopes for processes simulated from the various approximations. In all scenarios, the BDD and BDA envelopes for both mRNA and protein are closer to the true envelopes than the LNA with the truncated normal approximation modelling the skewness at low molecular numbers better than the symmetric LNA. The LNA improves as molecular numbers increase although consistently overestimates the variance for low numbers and will consequently be likely to miss switch points in the transcriptional profiles. The envelopes of the CLE again match very closely to the true process, however, since the transition densities are intractable, inference becomes much more challenging. This empirical validation reinforces the intuition that the BDA is more accurate at low molecular levels and may be preferable for inference to the more standard LNA, especially for systems of low molecular levels.

The inferential problem will be investigated in the following chapter and it will be shown that the LNA, although less accurate, has many advantages from an inferential viewpoint.

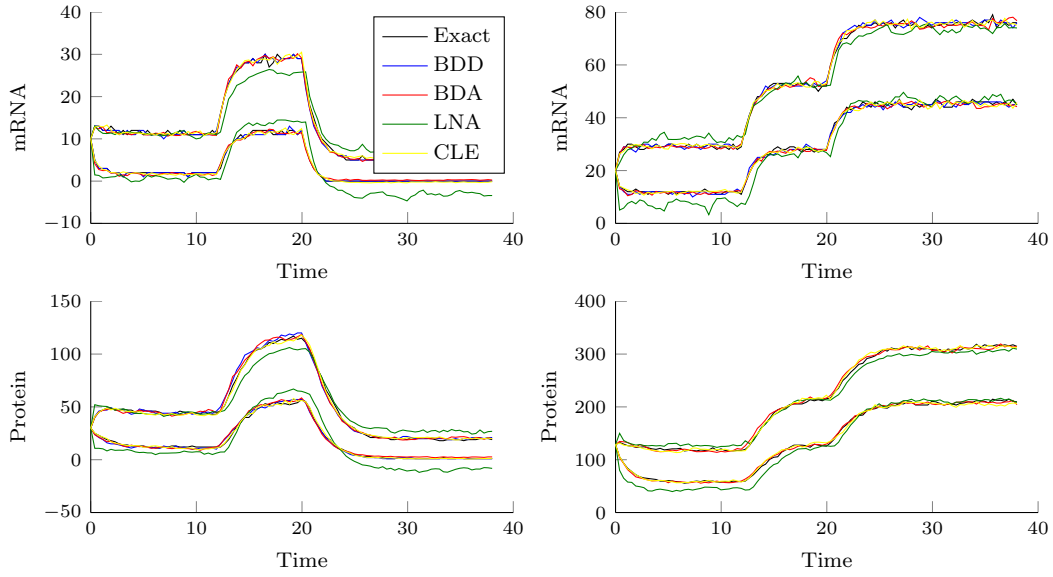


Figure 2.5: 95% pointwise confidence envelopes for simulated mRNA (top) and protein (bottom) processes under the true process (black), the BDD (blue), the truncated normal BDA (red), the LNA (green) and the CLE (yellow) for two different scenarios corresponding to different molecular abundances.

CHAPTER 3

INFERENCE FOR STATE SPACE MODELS OF STOCHASTIC GENE TRANSCRIPTION

Research; the curiosity to find the unknown to make it known.

Lailah Gifty Akita, Beautiful Quotes

3.1 State Space Representation

State space models provide a unifying framework for formulating each of the approximations described in the previous chapter. Formulating these approximations in a state space model enables one to make use of a number of techniques developed for inference on this class of models, in particular in the presence of an observation equation. The general state space model is given by two equations, the measurement or observation equation defining the observable data, Y, and the state equation, which defines the unobserved state of the system, X. Although, both X and Y may be multivariate, we make no distinction with boldface fonts, since this will instead be used to denote a sequence of observations described in the following.

The observation equation is given by,

$$\begin{aligned} Y(t) &= G(t)X(t) + \epsilon(t) \\ \epsilon(t) &\sim \mathcal{N}(0, \Sigma_\epsilon), \end{aligned} \tag{3.1}$$

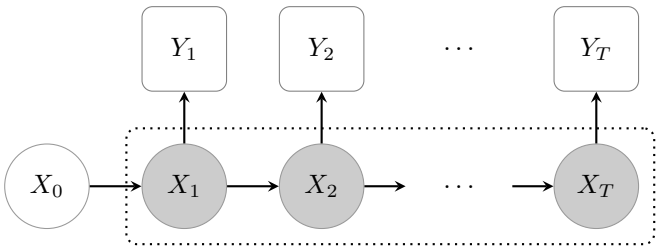


Figure 3.1: A pictorial representation of a general hidden Markov model for state space models. The observations $\mathbf{Y}$ are conditionally independent given the latent Markov process $\mathbf{X}$, with X_0 providing the prior information about the initial state.

for some (possibly time varying) matrix G. Although more general measurement processes can be modelled in the state space framework, we restrict ourselves to linear Gaussian measurement error since this provides a reasonable model for the measurement of light intensity levels in single cell gene expression experiments (see Section 4.1).

The state equation is given by,

$$\begin{aligned} X(t+\tau) &= F(t)X(t) + c(t) + \eta(t), \\ \eta(t) &\sim \mathcal{N}(\mu(t), \Sigma(t)), \end{aligned} \tag{3.2}$$

for matrices F, Σ and vectors c and μ. The state equation may be specified through any of the approximations outlined in the previous chapter. In what follows, it will be convenient to express equations (3.1) and (3.2) in terms of the following densities,

$$Y(t)|X(t) \sim g(y(t)|x(t), \theta), \tag{3.3}$$

$$X(t+\tau)|X(t) \sim h(x(t+\tau)|x(t), \theta), \tag{3.4}$$

where h and g denote the transition and observation densities respectively, with θ denoting the vector of (unknown) parameters.

Assuming data are observed at discrete time points, $t_1, ..., t_T$, we let $X_{t_i} := X(t_i)$, $Y_{t_i} := Y(t_i)$ and $y_{1:i}$ define the sequence of observations $y(t_1), ..., y(t_i)$ from time t_1 to t_i and $x_{1:i}$ define the corresponding sequence of latent states $x(t_1), ..., x(t_i)$. In addition, we will use the convention $\mathbf{y} := y_{1:T}$ to denote all the observed data points $y(t_1), .., y(t_T)$ and $\mathbf{x} := x_{1:T}$ to denote the corresponding latent states.

Assuming observations consist of only the data $\mathbf{y}$, the system can be expressed as a hidden Markov model (HMM), depicted in Figure 3.1, where the latent states $\mathbf{x}$ are unobserved. Since this chapter is predominantly interested in the inference problem

of estimating the posterior $f(\theta|\mathbf{y})$, we require the data likelihood given by,

$$f(\mathbf{y}|\theta) = \int_{\mathbf{x}} f(\mathbf{y}, \mathbf{x}|\theta) \ \mathrm{d}\mathbf{x}. \tag{3.5}$$

The integrand in the above, which we will refer to as the joint or extended likelihood, can be factorised into the following form, using the Markov property of the underlying process,

$$f(\mathbf{y}, \mathbf{x}|\theta) = \prod_{i=1}^{T} h(x(t_i)|x(t_{i-1}), \theta) g(y(t_i)|x(t_i), \theta). \tag{3.6}$$

It becomes clear from equation (3.6), why it is desirable to work directly with the transition densities. If one wanted to work with either the exact system or an approximation with an intractable transition density, one would have to further explicitly integrate each of the following,

$$h(x(t_i)|x(t_{i-1})) = \int_{t_{i-1}}^{t_i} \mathrm{SP}(\mathbf{x}) \ \mathrm{d}\mathbf{x}, \quad \forall i \in 1, ..., T,$$

where SP is the stochastic process of the state $\mathbf{x}$. Consequently, we restrict our attention to the LNA and BDA when considering inference for SRNs as tractable transition densities are available for each.

Despite the existence of analytical transition densities, it is not immediately obvious how one can obtain the data likelihood in equation (3.5). It is worthwhile noting, that the data likelihood can also be expressed through a simple application of Bayes rule by,

$$f(\mathbf{y}|\theta) = \frac{f(\mathbf{x}, \mathbf{y}|\theta)}{f(\mathbf{x}|\mathbf{y}, \theta)}.$$

However, it remains unclear how one obtains, or even samples from, the conditional density $f(\mathbf{x}|\mathbf{y}, \theta)$. We shall present the framework one can use to obtain both the data likelihood, $f(\mathbf{y}|\theta)$, and the latent conditional density, $f(\mathbf{x}|\mathbf{y}, \theta)$, for a completely general state space model in Section 3.2. Section 3.3 will then outline how this methodology can be implemented in a general inference framework.

The remainder of the chapter will be restricted to the LNA and BDA methodologies and the specific features of their associated state space formulation. In particular, Section 3.4 will discuss the computation of the likelihood under each of the LNA and BDA and compare features of these likelihood surfaces to the likelihood surface of the

exact process. Section 3.5 will return to our gene transcription model and describe how one can perform parameter inference on this system with the specification of all prior distributions given in Section 3.6. Section 3.7 discusses the problem of inferring the latent states under the BDA. Finally, the algorithm specification is given in Section 3.8 with an extensive simulation study presented in Section 3.9.

3.2 Filtering, Smoothing and Backward Sampling

Consider the latent states, $X_{t_1}, ..., X_{t_T}$ of the system depicted in Figure 3.1. Suppose we are interested in computing the conditional densities $f(x_{t_s}|y_{1:i})$. We shall see the relevance of these densities in the following, but first note that there are three distinct problems. Namely,

1. If $t_s = t_i$, estimating $f(x_{t_i}|y_{1:i})$ is the *filtering* problem with $f(x_{t_i}|y_{1:i})$, called the filtering density,

2. If $t_s < t_i$, estimating $f(x_{t_s}|y_{1:i})$ is the *smoothing* problem and,

3. If $t_s > t_i$, estimating $f(x_{t_s}|y_{1:i})$ is the *prediction* problem.

Since we are not interested in the behaviour of the latent states outside the observed time interval, we do not explicitly consider the prediction problem in this work. In contrast, both the filtering and smoothing problems have implications for inference, which we discuss here.

3.2.1 Filtering

Following Petris et al. (2009), for general state space models of the form given in equations (3.3) and (3.4), one can recursively obtain the filtering densities, $f(x_{t_i}|y_{1:i})$, for $i = 1, .., T$. Starting from the initial distribution $X_{t_0} \sim f(x_{t_0})$, the filtering densities are obtained in a recursive way through the following three steps.

1. Obtain the *one-step ahead state prediction* for $X_{t_i}|y_{1:i-1}$, given by the density,

$$f(x_{t_i}|y_{1:i-1}) = \int h(x_{t_i}|x_{t_{i-1}}) f(x_{t_{i-1}}|y_{1:i-1}) \ \mathrm{d}x_{t_{i-1}}, \tag{3.7}$$

   where $f(x_{t_{i-1}}|y_{1:i-1})$ is the previous filtered density.

2. Obtain the *one-step ahead observation prediction* for $Y_{t_i}|y_{1:i-1}$, given by the density,

$$f(y_{t_i}|y_{1:i-1}) = \int g(y_{t_i}|x_{t_i}) f(x_{t_i}|y_{1:i-1}) \ \mathrm{d}x_{t_i}. \tag{3.8}$$

3. Obtain the *filtering density* for $X_{t_i}|y_{1:i}$ through an application of Bayes rule to give,

$$f(x_{t_i}|y_{1:i}) = \frac{g(y_{t_i}|x_{t_i}) f(x_{t_i}|y_{1:i-1})}{f(y_{t_i}|y_{1:i-1})}. \tag{3.9}$$

Likelihood

Consequently, one can use the one-step ahead observation predictions given in equation (3.8) to obtain the likelihood. Explicitly, one can factorise the data likelihood into the following form,

$$f(\mathbf{y}|\theta) = f(y_{t_1}|\theta) \prod_{i=2}^{T} f(y_{t_i}|y_{1:i-1}, \theta), \tag{3.10}$$

given by the product of the one-step ahead observation prediction densities. Thus, through the filtering densities, one is able to obtain an analytical data likelihood provided there is an analytical expression for both the one-step ahead prediction densities and the filtering densities.

3.2.2 Smoothing and Backward Sampling

In contrast to the filtering problem, the smoothing problem instead considers the estimation of the latent states given *all* of the observed data. Starting from the final filtering density, $f(x_{t_T}|y_{1:T})$, the smoothing densities $f(x_{t_i}|y_{1:T})$, can be obtained in the following recursive way over $i = T-1, .., 1$,

1. Obtain the *backward transitions*, through Bayes rule,

$$f(x_{t_i}|x_{t_{i+1}}, y_{1:i}) = \frac{h(x_{t_{i+1}}|x_{t_i}) f(x_{t_i}|y_{1:i})}{f(x_{t_{i+1}}|y_{1:i})}, \tag{3.11}$$

   which involves both the one-step ahead state prediction density and filtering density of equations (3.7) and (3.9).

2. Obtain the *smoothing density* for $X_{t_i}|y_{1:T}$, given by,

$$f(x_{t_i}|y_{1:T}) = f(x_{t_i}|y_{1:i}) \int \frac{h(x_{t_{i+1}}|x_{t_i})}{f(x_{t_{i+1}}|y_{1:i})} f(x_{t_{i+1}}|y_{1:T}) \ \mathrm{d}x_{t_{i+1}}. \tag{3.12}$$

Latent States

Consequently, the conditional density over all the latent states, $f(\mathbf{x}|\mathbf{y}, \theta)$, can be factorised into the product of the backward transitions, since,

$$\begin{aligned} f(\mathbf{x}|\mathbf{y}, \theta) &= \prod_{i=1}^{T} f(x_{t_i}|x_{i+1:T}, y_{1:T}, \theta) \\ &= \prod_{i=1}^{T} f(x_{t_i}|x_{t_{i+1}}, y_{1:T}, \theta) \\ &= \prod_{i=1}^{T} f(x_{t_i}|x_{t_{i+1}}, y_{1:i}, \theta). \end{aligned}$$

Thus, one can obtain a sample of the latent states from the conditional density, $f(\mathbf{x}|\mathbf{y}, \theta)$, by sampling from each of the backward transition densities, if available. This procedure is known as the Forward-Filtering Backward Sampling (FFSB) algorithm.

3.3 Towards an Inference Framework

Section 3.2 described how one can obtain analytically the data likelihood for general state space models provided there exists an analytical form of the filtering and one-step ahead prediction densities. Consequently, we now consider two broad approaches to inference for state space models.

1. The data likelihood, $f(\mathbf{y}|\theta)$, is analytically available through the filtering recursions of Section 3.2.

2. The data likelihood is analytically unavailable, but the joint or extended likelihood, $f(\mathbf{y}, \mathbf{x}|\theta)$, has an analytical form.

In the first case, Bayesian inference about θ can be performed by sampling from the posterior $f(\theta|\mathbf{y})$. We shall see in Section 3.4.2, that since the measurement process

is assumed to be of linear Gaussian form, the LNA defines a linear Gaussian state space model and the data likelihood can be computed through the Kalman filter.

In the second case, Bayesian inference about θ can instead be performed on the extended posterior, $f(\theta, \mathbf{x}|\mathbf{y})$ where the latent states $\mathbf{x}$ are also estimated. In such scenarios, it will be assumed that inference about θ and $\mathbf{x}$ will be achieved through the following two-step Gibbs sampler,

1. Sample the parameter vector, θ, from the density $f(\theta|\mathbf{y}, \mathbf{x})$.
2. Sample the latent states, $\mathbf{x}$, from the conditional density $f(\mathbf{x}|\mathbf{y}, \theta)$.

3.4 The Likelihood

In this section, we give explicit formulae for the likelihood under each of the exact process, the LNA and the BDA and compare the likelihood surfaces of each.

3.4.1 Exact Likelihood

Recall that the exact likelihood is given by,

$$\begin{aligned} f(\mathbf{y}|\theta) &= \int_{\mathbf{x}} f(\mathbf{y}, \mathbf{x}|\theta) \ \mathrm{d}\mathbf{x}, \\ &= \int_{\mathbf{x}} g(\mathbf{y}|\mathbf{x}) \prod_{i=1}^{n} h_{j_i}(\mathbf{x}(t_i), \theta_j) \ \exp\left(-\sum_{i=0}^{n} h_0(\mathbf{x}(t_i), \theta)[t_{i+1} - t_i] \right) \ \mathrm{d}\mathbf{x}, \quad (3.13) \end{aligned}$$

where h_j is the hazard vector, n is the number of reactions that take place, $j_1, ..., j_n$ is the sequence of reaction types and $t_1, .., t_n$ are the associated timings of each reaction. Methods for inference of this likelihood were reviewed in Section 2.2 and are not considered any further. Instead we use (3.13), or specifically, the integrand of (3.13) to compare the correlation structures within the parameter vector θ for the LNA and BDA.

3.4.2 LNA Likelihood

Under the LNA, the likelihood is given by,

$$f(\mathbf{y}|\theta) = \int_{\mathbf{x}} f(\mathbf{y}, \mathbf{x}|\theta) \ \mathrm{d}\mathbf{x},$$

$$= \int_{\mathbf{x}} \prod_{i=1}^{T} h(x(t_i)|x(t_{i-1}), \theta) g(y(t_i)|x(t_i), \theta) \ \mathrm{d}\mathbf{x}, \tag{3.14}$$

where h is given in equation (2.30) and g is as in (3.3). Specifically, under the LNA, the state equation (3.2) becomes a linear Gaussian state equation of the form (Finkenstädt et al., 2013),

$$X(t_{i+1}) = F(t_i)X(t_i) + c(t_i) + \eta(t_i), \tag{3.15}$$

$$\eta(t_i) \sim N(0, \Sigma(t_i)). \tag{3.16}$$

Kalman Methodology

In order to obtain the data likelihood under the LNA, we need to perform the filtering recursions described in Section 3.2.1. One can explicitly evaluate each of the required densities, moreover, noting that under a linear Gaussian model, the vector $(X_{t_1}, ..., X_{t_T}, Y_{t_1}, ..., Y_{t_T})$ has a multivariate Gaussian distribution, it follows that all the conditional and marginal densities required in the filtering algorithm will be Gaussian and it suffices to simply evaluate the mean and variance (Petris et al., 2009). Filtering on the linear Gaussian state space model is attributed to Kalman (1960) and consequently has been termed the Kalman Filter.

To start the recursions, one must first assign a prior distribution to the initial latent states, $X(t_0) \sim N(a_0, Q_0)$. The recursions then take the following form. For time $i = 1, ..., T$,

1. The *one-step ahead state prediction* density is given by, $X(t_i)|y_{1:t_{i-1}} \sim N(b_i, R_i)$, where the predictive equations are given by,

$$b_i = F(t_{i-1})a_{i-1} + c(t_{i-1}),$$
$$R_i = F(t_{i-1})Q_{i-1}F(t_{i-1})^T + \Sigma(t_{i-1}).$$

2. The *one-step ahead observation prediction* density is given by, $Y(t_i)|y_{1:t_{i-1}} \sim N(G(t_i)b_i, S_i)$ where,

$$S_i = G(t_i)R_{i-1}G(t_i)^T + \Sigma_\epsilon. \tag{3.17}$$

3. Finally, the *filtering density* is given by $X(t_i)|y_{1:t_i} \sim N(a_i, Q_i)$, where the

updating equations satisfy,

$$\begin{aligned} a_i &= b_i + R_i G(t_i)^T S_i^{-1} \left(y(t_i) - G(t_i) b_i \right) \\ Q_i &= R_i - R_i G(t_i)^T S_i^{-1} G(t_i) R_i. \end{aligned} \tag{3.18}$$

The likelihood for the data $\mathbf{y} := y_{1:T}$ is then given by the product of the one-step ahead observation predictive distributions,

$$\begin{aligned} f(\mathbf{y}|\theta) =& f(y_1|\theta) \prod_{i=2}^{T} f(y_{t_i}|y_{1:t_{i-1}}, \theta) \\ =& \tfrac{1}{(2\pi)^{T/2}|S_1|} \exp\left(-\tfrac{1}{2} \left(y(t_1) - G(t_1) b_1 \right)^T S_1^{-1} \left(y(t_1) - G(t_1) b_1 \right) \right) \times ... \\ & ... \times \prod_{i=2}^{T} \tfrac{1}{(2\pi)^{n/2}|S_i|} \exp\left(-\tfrac{1}{2} \left(y(t_i) - G(t_i) b_i \right)^T S_i^{-1} \left(y(t_i) - G(t_i) b_i \right) \right). \end{aligned}$$

In order to use the above methodology, one needs to specify a starting value for the recursions, a_0 and Q_0. These starting values may depend upon the parameter vector θ. In particular, Komorowski et al. (2009) let $a_0 := x_0$ be the initial value of the latent states and treat them as additional parameters to be estimated and $Q_0 := \Sigma_0$ is chosen to satisfy the following equation,

$$0 = J_0 \Sigma_0 + \Sigma_0 J_0^T + B_0 B_0^T,$$

where $J_0 := J(\phi(0))$ and $B_0 := B(\phi(0))$. This ensures that the initial covariance matrix is set to be the covariance of the system at time $t = 0$ if the system was initialised at a steady state.

The restarting variant of the LNA (Fearnhead et al., 2014) uses the filtering distribution calculated in (3.18), to reset the ODEs, ϕ, used in the LNA transition densities. Specifically, the ODE is recalculated at each time point subject to the condition, $\phi(t_{i-1}) = a_{i-1}$, where $\phi(t_{i-1})$ is the ODE evaluated at time point t_{i-1}. Note that under this Gaussian framework, a_{i-1} is the best linear unbiased predictor of $\mathbf{X}(t_{i-1})$ and is often denoted by $\hat{\mathbf{X}}(t_{i-1})$. The restarting LNA method of Fearnhead et al. (2014) is essential for non-linear systems as it reduces the impact of the initial value. For linear systems, the differences between the restarting and non-restarting methods is reduced (see Section 3.4.4). In addition, due to the recursive nature of the restarting method, the implementation can become considerably slower and we consequently focus our attention on the non-restarting version for this (piecewise-linear) application of gene transcription.

3.4.3 BDA Likelihood

Under the BDA, the likelihood is given by,

$$\begin{aligned} f(\mathbf{y}|\theta) &= \int_{\mathbf{x}} f(\mathbf{y}, \mathbf{x}|\theta) \ \mathrm{d}\mathbf{x}, \\ &= \int_{\mathbf{x}} \prod_{i=1}^{T} h(x(t_i)|x(t_{i-1}), \theta) g(y(t_i)|x(t_i), \theta) \ \mathrm{d}\mathbf{x}, \end{aligned} \tag{3.19}$$

where g is as in (3.3) and $h(x(t_i)|x(t_{i-1})) := h_m(m(t_i)|m(t_{i-1}))h_p(p(t_i)|p(t_{i-1}), m^*(t_i))$, such that,

$$\begin{aligned} h_m &\sim N_T\left(\lambda^m(t_i) + \pi^m(t_i)m(t_{i-1}), \lambda^m(t_i) + \pi^m(t_i)(1 - \pi^m(t_i))m(t_{i-1})\right), \\ h_p &\sim N_T\left(\lambda^p(t_i) + \pi^p(t_i)p(t_{i-1}), \lambda^p(t_i) + \pi^p(t_i)(1 - \pi^p(t_i))p(t_{i-1})\right), \end{aligned}$$

where $\lambda := (\lambda^m, \lambda^p)^T$ and $\pi := (\pi^m, \pi^p)^T$ satisfy the system of ODEs (2.38)-(2.39).

Although, this likelihood is defined through Gaussian densities (albeit truncated to the positive real line), unlike the LNA, it does not define a linear state space model. This is due to the non-linear dependence upon the latent states in the variance term of the transition density. Consequently, one cannot use the Kalman methodology to evaluate (3.19) explicitly.

As discussed in Section 3.3, one can instead view $f(\mathbf{x}, \mathbf{y}|\theta)$ as the joint likelihood of the system and the problem reduces to a simulation framework, where one samples the latent states from the conditional density $f(\mathbf{x}|\mathbf{y}, \theta)$. One way in which these samples can be obtained is through the forward filtering backward sampling algorithm outlined in Section 3.2.2. However, this is only possible if one can evaluate the backward transitions, $f(x_{t_i}|x_{t_{i+1}}, y_{1:i})$, which under the BDA, are not available analytically. There are a number of other approaches one can use to sample the latent states, a selection of which are presented below.

Sampling Latent States

The first method is to estimate the conditional density $f(\mathbf{x}|\mathbf{y}, \theta)$ through an MCMC sampler. This could be achieved through a Gibbs sampler, where at each iteration of the MCMC algorithm, one would sample each $x(t_i)$ from the full conditional, $f(x(t_i)|\mathbf{x}_{1:t_{i-1}}, \mathbf{x}_{t_{i+1}:T}, \mathbf{y}, \theta)$. In non-linear or non-Gaussian frameworks, the density $f(x(t_i)|\mathbf{x}_{1:t_{i-1}}, \mathbf{x}_{t_{i+1}:T}, \mathbf{y}, \theta)$ may only be known up to a normalising constant and a

Metropolis within Gibbs algorithm may be used to sample from these full conditionals.

In general, due to the large number of latent states, this approach could result in very slow mixing Markov chains. Moreover, latent states are likely to be highly correlated and this approach will be inefficient for sampling from the full posterior distribution and one could instead update blocks of latent states. Rather than a Metropolis step to propose the block of latent states, one could instead use an independence sampler of the form given in Algorithm 1.

Algorithm 1 Updating the hidden states of a HMM via an independence sampler

1: Sample two observation points $X(t_{k-1})$ and $X(t_{k+m+1})$ at random, where m is the distance between them.
2: Propose new values for $\mathbf{X}^*_{t_k:t_m} := X^*(t_k), ..., X^*(t_m)$, from a proposal distribution q.
3: Accept this new path with probability α given by,

$$\alpha = \min\left(1, \frac{q(\mathbf{x}_{t_k:t_m}) f(\mathbf{y}, \mathbf{x}^*|\theta)}{q(\mathbf{x}^*_{t_k:t_m}) f(\mathbf{y}, \mathbf{x}|\theta)}\right) \tag{3.20}$$

Consequently, given two endpoints, $X(t_{k-1})$ and $X(t_{k+m+1})$, a proposal distribution, q, is required to propose $\mathbf{X}^*_{t_k:t_m}$. Specifically, we consider the Laplace approximation of $f(\mathbf{y}, \mathbf{x}_{k:m}|\theta)$, where the proposal distribution is given by a multivariate normal distribution with mean given by the mode, $\tilde{\mathbf{x}}_{k:m} := \text{argmax}_{\mathbf{x}} f(\mathbf{y}, \mathbf{x}_{k:m}|\theta)$, obtained by a small number of Newton steps, and covariance matrix given by the negative inverse Hessian, $\Sigma := -H^{-1}$, where $H_{km} := \left(\frac{\partial^2 f}{\partial x_k \partial x_m}\right)$. This approach has previously been implemented within a bridging framework (Elerian et al., 2001) that Heron et al. (2007) applied to stochastic kinetic models. However, this approach has a non-trivial computational cost due to the high dimensional posterior. In practice, the maximisation of the Laplace approximation is computed using at most three Newton steps to avoid high computational cost. To avoid a low acceptance rate, we update only short sections of the trajectory at each iteration, which can result in a slow convergence rate of the Markov chains.

In scenarios of low molecular numbers, the Laplace/independence sampler used with the BDA often failed. This is unsurprising since the true latent states will lie on or close to the state boundary (note that the support of the latent states is the positive real line) and thus the mode of the normal approximation to the posterior will lie close to the boundary, resulting in a non-positive semi-definite Hessian preventing

its use in the proposal distribution for latent states. Moreover, in scenarios of larger molecular numbers, although valid, the Laplace method is extremely slow owing to the optimisation step used to compute the proposals. We therefore generally regard this as an unfeasible method to use.

Particle Filter

A class of methods that have become popular in recent years for sampling hidden states in a hidden Markov model are particle filters. This is a sequential Monte Carlo (SMC) approach based on forward simulations to sequentially approximate the conditional densities, $f(x_{1:t_i}|y_{1:t_i})$ and can be applied to very general state space models that are not necessarily linear or Gaussian. Specifically, we shall consider the sequential importance sampling (SIS) algorithm of Doucet et al. (2000). Following the motivation of SMC presented in Petris et al. (2009), the basic principle of the SMC methodology is found in importance sampling. Suppose one is interested in evaluating the expected value,

$$\mathbb{E}_f[g(X)] = \int_x g(x)f(x)\ \mathrm{d}x. \tag{3.21}$$

Letting q be some importance density, equation (3.21) can be rewritten in the form,

$$\begin{aligned}\mathbb{E}_f[g(X)] &= \int_x g(x)\frac{f(x)}{q(x)}q(x)\ \mathrm{d}x \\ &= \mathbb{E}_q[w(X)],\end{aligned} \tag{3.22}$$

where w is the importance function given by $g(x)f(x)/q(x)$ and q is the importance density. Consequently, the main idea behind importance sampling is that one can generate a sample from q and compute,

$$\tfrac{1}{N_p}\sum_{j}^{N_p} w(x^{(j)}),$$

to get an approximation of equation (3.22). Therefore setting $g(x) := \delta_x$ to be the delta function centred at x, the unknown density, f, can be approximated by the discrete approximation,

$$f \approx f^{N_p} = \tfrac{1}{N_p}\sum_{j=1}^{N_p} w(x^{(j)}).$$

Thus, to estimate an unknown density, one requires a sample $\{x^{(j)}, w^{(j)} : j = 1, ..., N_p\}$, where $x^{(j)}$ are samples or particles from some importance density and $w^{(j)}$ are the associated importance weights. Consequently, the particles $x^{(j)}$ form a weighted sample from the density f.

In order to estimate the conditional density $f(\mathbf{x}|\mathbf{y}, \theta)$ through an importance sampling approach, one needs to construct an efficient proposal distribution. Typically, in state space models, the dimension of $f(\mathbf{x}|\mathbf{y}, \theta)$ is very large and it is difficult to obtain samples with reasonable weight. The sequential importance sampling methodology alleviates this issue by exploiting the recursive properties of the model. Explicitly, the conditional densities $f(x_{1:t_i}|y_{1:t_i})$ satisfy,

$$f(x_{1:t_{i+1}}|y_{1:t_{i+1}}) \propto g(y_{t_{i+1}}|x_{t_{i+1}})h(x_{t_{i+1}}|x_{t_i})f(x_{1:t_i}|y_{1:t_i}),$$

where h and g are the state and observation densities as given in (3.19). This means that both the importance density and importance weights can be defined recursively,

$$\begin{aligned} q(x_{1:t_{i+1}}|y_{1:t_{i+1}}) &\propto q(x_{t_{i+1}}|y_{t_{i+1}}, x_{t_i})q(x_{1:t_i}|y_{1:t_i}), \\ w_{t_{i+1}}^{(j)} &= \frac{f(x_{1:t_{i+1}}|y_{1:t_{i+1}})}{q(x_{1:t_{i+1}}|y_{1:t_{i+1}})} \\ &\propto w_t^{(j)} \frac{g(y_{t_{i+1}}|x_{t_{i+1}})h(x_{t_{i+1}}|x_{t_i})}{q(x_{t_{i+1}}|y_{t_{i+1}}, x_{t_i})}. \end{aligned}$$

Consequently, one can sequentially draw samples from $f(x_{1:t_i}|y_{1:t_i})$, to end with a weighted sample $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N_p)}$ from the full conditional density $f^{N_p}(x_{1:T}|y_{1:T})$. In practice, due to the sequential nature of this sampling procedure, it can lead to a *weight degeneracy* problem, where only a small number of samples have a significant weight by the end of the recursions. For example, as depicted in Figure 3.2a), all initial samples of x_1 have reasonable weight but sampling forward in time results in many samples of x_T having very low weight and being far from the true path. To overcome this issue, one can use a resampling procedure where the particle approximation $\{x_{1:t_i}^{(j)}, w_{t_i}^{(j)}\}$ is transformed into an equally weighted sample by sampling with replacement (Figure 3.2b)). The procedure is then termed the sequential importance resampling (SIR) algorithm and is given in Algorithm 2.

Algorithm 2 SIR algorithm

1: At time $t = t_1$, sample N_p particles from initial distribution $X_{t_1} \sim h(x_{t_1}|x_{t_0})$ to obtain a sample $x_{t_1}^{(1)}, ..., x_{t_1}^{(N_p)}$. Compute the weights and normalise, where,

$$w_{t_1}^{(j)} \propto \frac{g\left(y_{t_1}|x_{t_1}^{(j)}\right) h\left(x_{t_1}^{(j)}|x_{t_0}\right)}{q\left(x_{t_1}^{(j)}|y_{t_1}\right)}.$$

2: For $i = 2, ..., T$,

a) For $j = 1, ..., N_p$,
sample $X_{t_i}^{(j)} \sim q\left(.|x_{1:t_{i-1}}^{(j)}, y_{1:t_i}\right)$ to obtain N_p paths $x_{1:t_i}^{(1)}, ..., x_{1:t_i}^{(N_p)}$.

b) Calculate the incremental importance weights,

$$w_{t_i}^{(j)} \propto w_{t_{i-1}}^{(j)} \frac{g\left(y_{t_i}|x_{t_i}^{(j)}\right) h\left(x_{t_i}^{(j)}|x_{t_{i-1}}^{(j)}\right)}{q\left(x_{t_i}^{(j)}|x_{1:t_{i-1}}^{(j)}, y_{1:t_i}\right)}.$$

c) Normalise weights,

$$W_{t_i}^{(j)} = \frac{w_{t_i}^{(j)}}{\sum_{k=1}^{N_p} w_{t_i}^{(k)}}.$$

d) Calculate the estimated effective sample size,

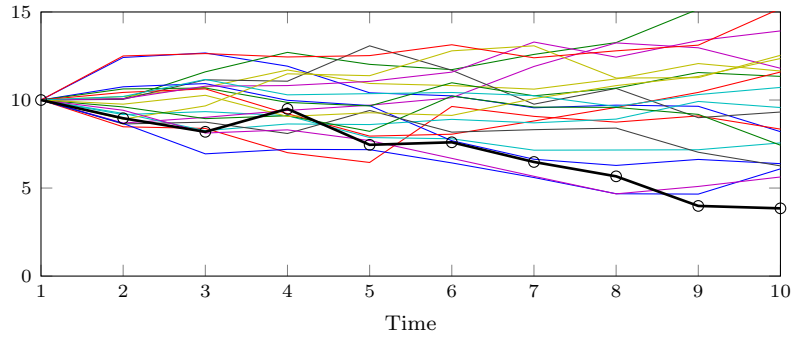
$$\hat{N}_{\text{eff}} = 1/\sum_{j=1}^{N_p} \left(W_{t_i}^{(j)}\right)^2.$$

If $\hat{N}_{\text{eff}} < \hat{N}_{\text{Thres}}$, resample the paths with weights, $W_{t_i}^{(j)}$, and set $W_{t_i}^{(j)} = \frac{1}{N_p}$ for $j = 1, ..., N_p$.

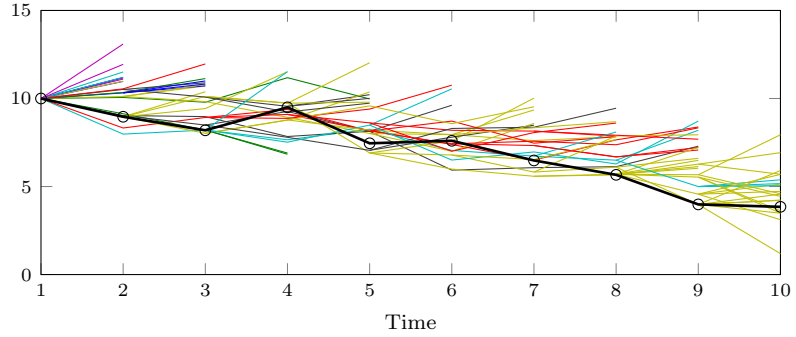
3: Sample from $1, ..., N_p$ with weights $W_T^{(1)}, ..., W_T^{(N_p)}$ to obtain a sample of the full latent path $x_{1:T}^{(B)}$ consistent with the data $\mathbf{y}$. Define the ancestral lineage, $B = (B_1, ..., B_T)$ of the sample as the sequence of indices of the particle parents.

We define the ancestral lineage, $B = (B_1, ..., B_T)$ of the sample as the sequence of indices of the particle parents. This ancestral lineage becomes important when incorporating the particle filter into a conditional SMC framework as described in Section 3.7.

A key technicality of the particle filter is the resampling step 2d) of Algorithm 2. This is used to ensure we can efficiently sample paths compatible with the data,



(a)



(b)

Figure 3.2: Illustration of weight degeneracy in a) where by time point 10, very few paths have reasonable weight conditional on the observed data shown in black. b) gives an illustration of particle degeneracy where the paths are resampled to maintain reasonable weight but there are very few independent paths remaining by time point 10.

although it may result in *particle degeneracy* where only a small number of independent paths are obtained as depicted in Figure 3.2b). Thus, we need to sample enough particles to ensure that there is a sufficient number of independent paths at the end of the algorithm.

In addition, in order to implement the particle filter, one needs to define a proposal distribution. This proposal distribution is used to draw samples of the latent process, x_{t_i}, conditional on the observations up to time t_i and the latent process up to time t_{i-1}. Since this algorithm is an importance sampler, the proposal distribution, q, will be optimal if it is equal to the density,

$$q(x_{t_i}|x_{1:t_{i-1}}, y_{1:t_i}) = f(x_{t_i}|x_{1:t_{i-1}}, y_{1:t_i}).$$

If this density is not analytically available, q should otherwise be chosen as a good approximation to $f(x_{t_i}|x_{1:t_{i-1}}, y_{1:t_i})$ with slightly heavier tails.

Example. *Implementation under the BDA.* Consider the gene transcription model (2.3) - (2.4) under the BDA. Recall that this approximation follows the non-linear state space model,

$$\begin{aligned} Y_{t_i} &\sim N(\begin{pmatrix} 0 & \kappa \end{pmatrix} X_{t_i}, \sigma_\epsilon^2), \\ X_{t_{i+1}} &\sim N_T(\mu_{t_{i+1}}, \sigma_{t_{i+1}}^2), \end{aligned}$$

where $X_{t_{i+1}} := (M_{t_{i+1}}, P_{t_{i+1}})^T$, and N_T is a truncated Gaussian density with,

$$\begin{aligned} \mu_{t_{i+1}} &= \lambda_{t_i} + \pi_{t_i} X_{t_i}^T, \\ \sigma_{t_{i+1}}^2 &= \lambda_{t_i} + \pi_{t_i}(1 - \pi_{t_i}) X_{t_i}^T), \end{aligned}$$

with the vectors $\lambda_{t_i} := (\lambda_{t_i}^m, \lambda_{t_i}^p)^T$ and $\pi_{t_i} := (\pi_{t_i}^m, \pi_{t_i}^p)^T$ satisfying the ODE system (2.38)-(2.39). Consequently, under the BDA, the joint transition density $h(x_{t_{i+1}}|x_{t_i})$ can be decomposed into,

$$h(m_{t_{i+1}}, p_{t_{i+1}}|m_{t_i}, p_{t_i}) = h_m(m_{t_{i+1}}|m_{t_i}) h_p(p_{t_{i+1}}|m_{t_{i+1}}, p_{t_i}),$$

where h_m is the marginal transition for the mRNA process and h_p is the BDA marginal transition density for the protein process. Thus, the joint filtering density can also be decomposed as follows,

$$\begin{aligned} f(m_{t_i}, p_{t_i}|m_{1:t_{i-1}}, p_{1:t_{i-1}}, y_{1:t_i}) &= f(p_{t_i}|m_{1:t_{i-1}}, m_{t_i}, p_{1:t_{i-1}}, y_{1:t_i}) \times \\ &\quad \times f(m_{t_i}|m_{1:t_{i-1}}, p_{1:t_{i-1}}, y_{1:t_i}) \\ &= f(p_{t_i}|m_{t_i}, p_{t_{i-1}}, y_{t_i}) h_m(m_{t_i}|m_{t_{i-1}}). \end{aligned}$$

Letting g denote the observation density, one can propose $X_{t_i} := (M_{t_i}, P_{t_i})$ in two steps.

1. Propose M_{t_i} from the transition density $h_m(m_{t_i}|m_{t_{i-1}})$, which will have importance weight proportional to 1.

2. Secondly, propose P_{t_i} from the proposal $q(p_{t_i}|m_{t_i}, p_{t_{i-1}}, y_{t_i})$ given in (3.23) and derived below.

   In order to construct a reasonable proposal distribution, q, we note that,

   $$f(p_{t_i}|m_{t_i}, p_{t_{i-1}}, y_{t_i}) \propto h_p(p_{t_i}|p_{t_{i-1}}, m_{t_i}) g(y_{t_i}|p_{t_i}),$$

   where h_p is the truncated Gaussian transition density for protein with mean

$\mu^p_{t_i} := \lambda^p + p_{t_{i-1}}\pi^p$ and variance $\sigma^{p^2}_{t_i} := \lambda^p + p_{t_{i-1}}\pi^p(1-\pi^p)$. Consequently, under the corresponding (non-truncated) normal approximation to h_p, (i.e. $h^*_p \sim N(\mu^p_{t_i}, \sigma^{p^2}_{t_i})$),

$$f(p_{t_i}|m_{t_i}, p_{t_{i-1}}, y_{t_i}) \overset{\text{approx}}{\propto} h^*_p(p_{t_i}|p_{t_{i-1}}, m_{t_i})g(y_{t_i}|p_{t_i}),$$

thus enabling the construction of the following proposal distribution for $P_{t_i}|P_{t_{i-1}}, M_{t_i}, Y_{t_i}$,

$$\begin{aligned} P_{t_i}|P_{t_{i-1}}, M_{t_i}, Y_{t_i} &\sim N(\mu^*_{t_i}, \sigma^{*^2}_{t_i}) \\ \sigma^{*^2}_{t_i} = \left(\frac{1}{\sigma^{p^2}_{t_i}} + \frac{\kappa^2}{\sigma^2_\epsilon}\right)^{-1}, \quad \mu^*_{t_i} &= \sigma^{*^2}_{t_i}\left(\frac{\mu_{t_i}}{\sigma^{p^2}_{t_i}} + \frac{\kappa y_{t_i}}{\sigma^2_\epsilon}\right). \end{aligned} \tag{3.23}$$

As with the LNA, one requires the initial state $X_{t_0} := (M_{t_0}, P_{t_0})^T$ to initialise the algorithm and are treated as additional model parameters.

Although, the particle filter comes at its own computational cost when incorporating it within the overall Bayesian framework for inference about the parameters, θ (Section 3.7), as a method for obtaining samples from the BDA conditional density $f(\mathbf{x}|\mathbf{y}, \theta)$, it is considerably faster and more efficient than the MCMC methods discussed in this section. Therefore, we shall restrict our attention to this methodology for estimating the conditional density $f(\mathbf{x}|\mathbf{y}, \theta)$, under the BDA.

Pseudo-Marginal Approach

The above approaches all rely on sampling the latent states of the system so that one can compute the joint likelihood $f(\mathbf{x}, \mathbf{y}|\theta)$ and consequently target the extended posterior $f(\theta, \mathbf{x}|\mathbf{y})$. Since our main interest is in the posterior $f(\theta|\mathbf{y})$, an alternative approach under the BDA is the pseudo-marginal method of Beaumont (2003) and Andrieu and Roberts (2009). In this framework, the likelihood $f(\mathbf{y}|\theta)$ is approximated by an unbiased Monte Carlo estimate $\hat{f}(\mathbf{y}|\theta)$. Explicitly,

$$\begin{aligned} \hat{f}(\mathbf{y}|\theta) &= \hat{f}(y_{t_1}|\theta)\prod_{i=1}^{T-1}\hat{f}(y_{t_{i+1}}|y_{t_1}, ..., y_{t_i}, \theta) \\ &= \prod_{i=1}^{T-1}\tfrac{1}{N_p}\sum_{j=1}^{N_p}\omega^{(j)}_{t_{i+1}}, \end{aligned}$$

where $\omega_{t_{i+1}}^{(j)} := f(y_{t_{i+1}}|x_{t_{i+1}}^{(j)}, \theta)$ and $x_{t_{i+1}}^{(j)}$ are the samples from $f(\mathbf{x}|\mathbf{y}, \theta)$. Thus, given any of the methods described above for obtaining samples of the latent states from the full conditional density, one can obtain an estimate of the data likelihood.

3.4.4 Likelihood Comparison

Finally we compare some features of the corresponding likelihood function for each model. Data were generated from the exact process via a stochastic simulation algorithm (Gillespie, 1977) based on equation (2.5). Sampling these data at discrete time points enabled us to evaluate the likelihood under the two approximations given by the LNA and BDA. In general (plots not shown), we confirmed that the higher the sampling frequency, the gradient of the likelihood became steeper about the maxima.

In order to compare the features of the likelihood, we first consider the case when latent states are observed and the likelihood is given by the joint density, $f(\mathbf{y}, \mathbf{x}|\theta)$. The bivariate extended likelihood surfaces are then presented in Figure 3.3 for the exact process, Figure 3.4 for the LNA and Figure 3.5 for the BDA. It can be seen that under the two approximations, significantly more correlation structure appears between the different parameters. In particular, we see a high correlation between δ_m and β and also between α and δ_p. This is unsurprising, since it is the trade-off between the birth and death rates of the two species.

However, in practice, the latent states are not observed and the data likelihood is given by $f(\mathbf{y}|\theta)$. Since this is only available analytically for the LNA, we have only shown the likelihood surfaces in Figures 3.6 and 3.7 for the restarting and non-restarting LNA respectively. There is very little difference between the LNA and restarting LNA likelihood surfaces but notably, there is a great deal more structure in the surfaces compared to the case when the latent states are observed. Specifically, there is a strong correlation between most of the parameters with only the measurement error, σ_ϵ, appearing to be independent. It is likely that the likelihood surface under the BDA for unobserved latent states will also show greater structure and consequently there is a need to design an efficient MCMC sampler to explore the highly correlated posterior density.

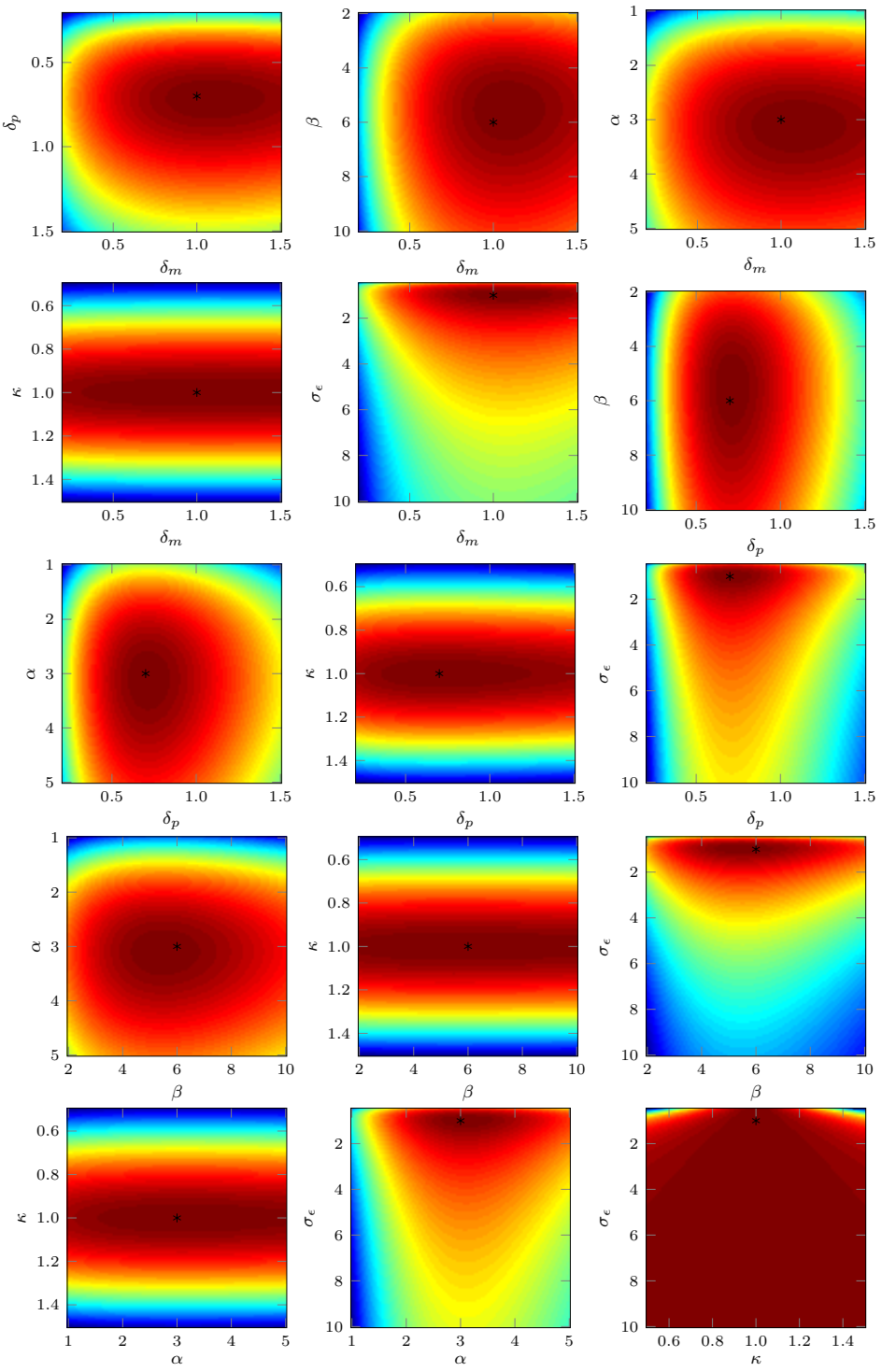


Figure 3.3: Bivariate likelihood transects under the exact joint likelihood. True parameter values are shown by the black points. Data were generated via an exact stochastic simulation algorithm. Latent states were fixed at their true values with observations taken over continuous time.

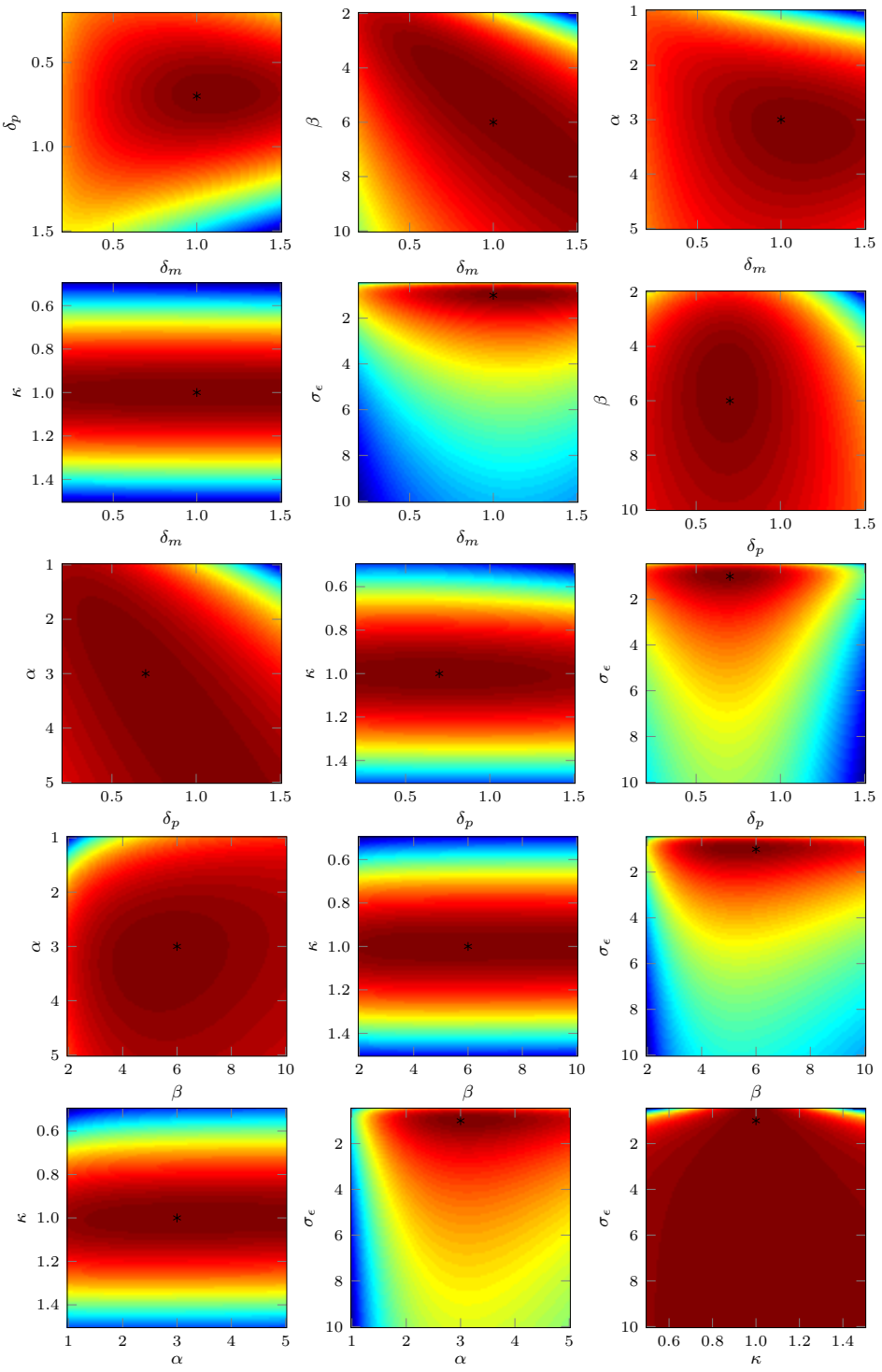


Figure 3.4: Bivariate likelihood transects under the LNA joint likelihood. True parameter values are shown by the black points. Data were generated via an exact stochastic simulation algorithm and sampled at 100 discrete time points. Latent states were fixed at their true values.

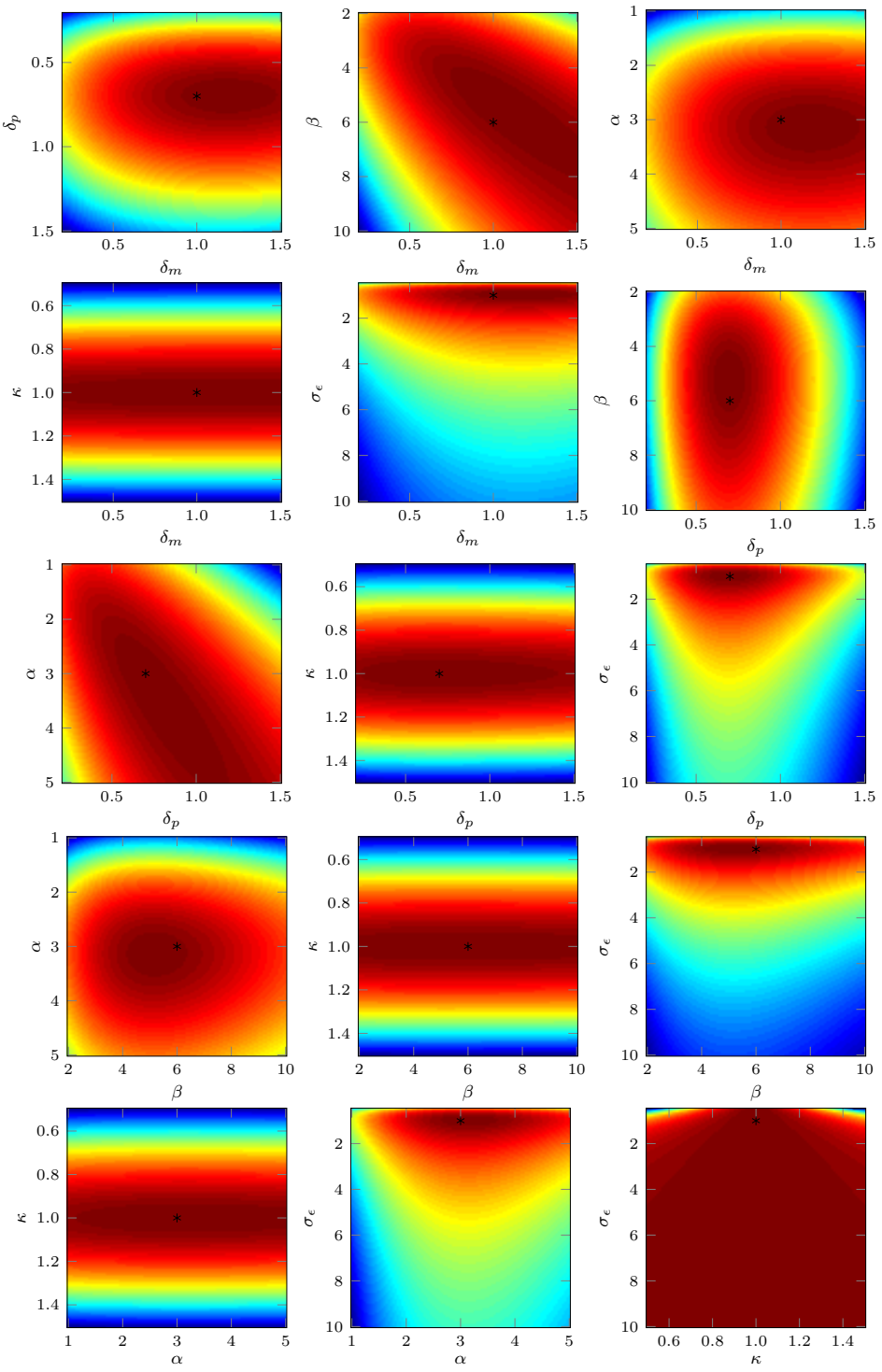


Figure 3.5: Bivariate likelihood transects under the BDA joint likelihood. True parameter values are shown by the black points. Data were generated via an exact stochastic simulation algorithm and sampled at 100 discrete time points. Latent states were fixed at their true values.

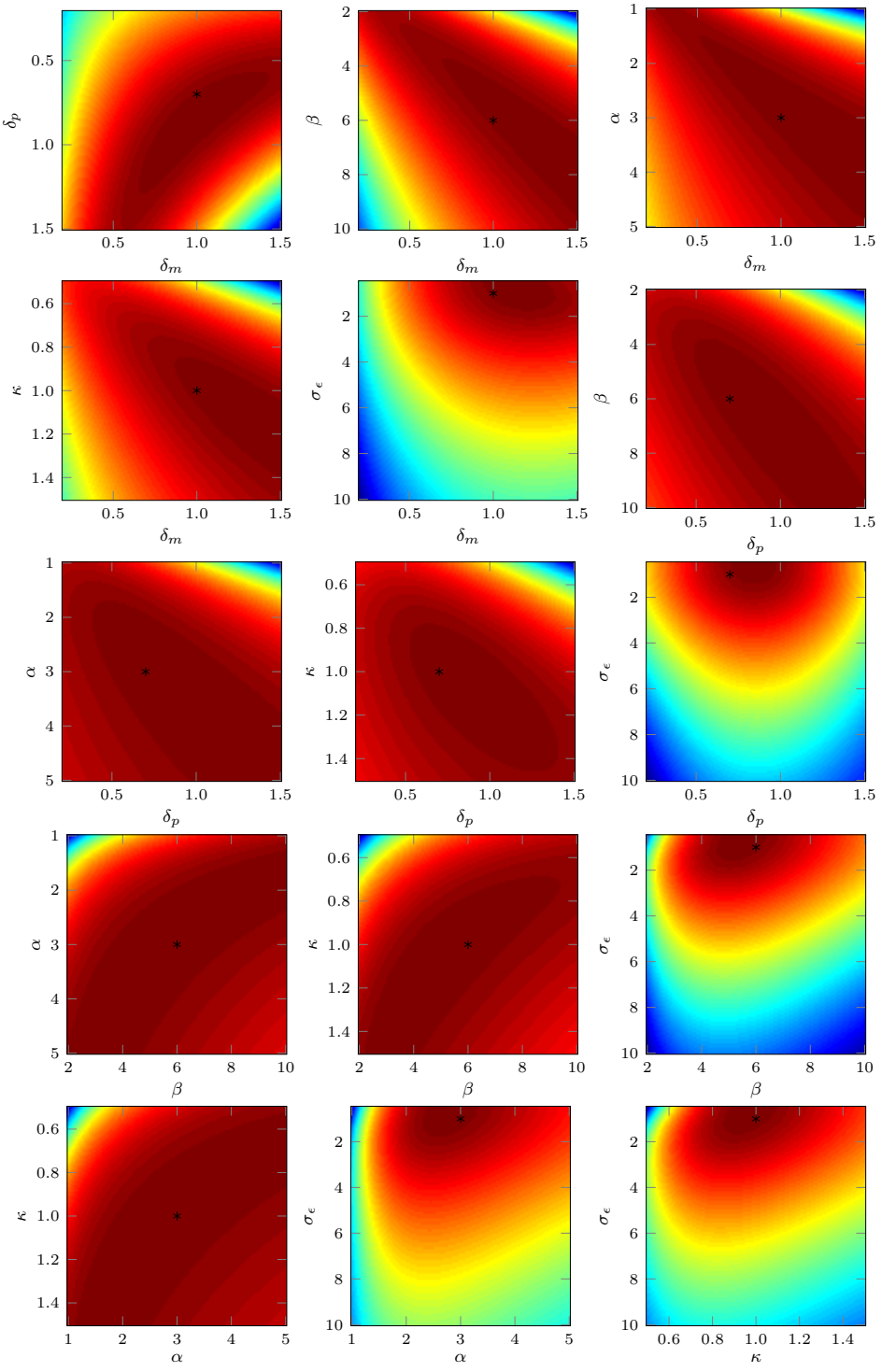


Figure 3.6: Bivariate likelihood transects under the restarting LNA data likelihood. True parameter values are shown by the black points. Data were generated via an exact stochastic simulation algorithm and sampled at 100 discrete timepoints.

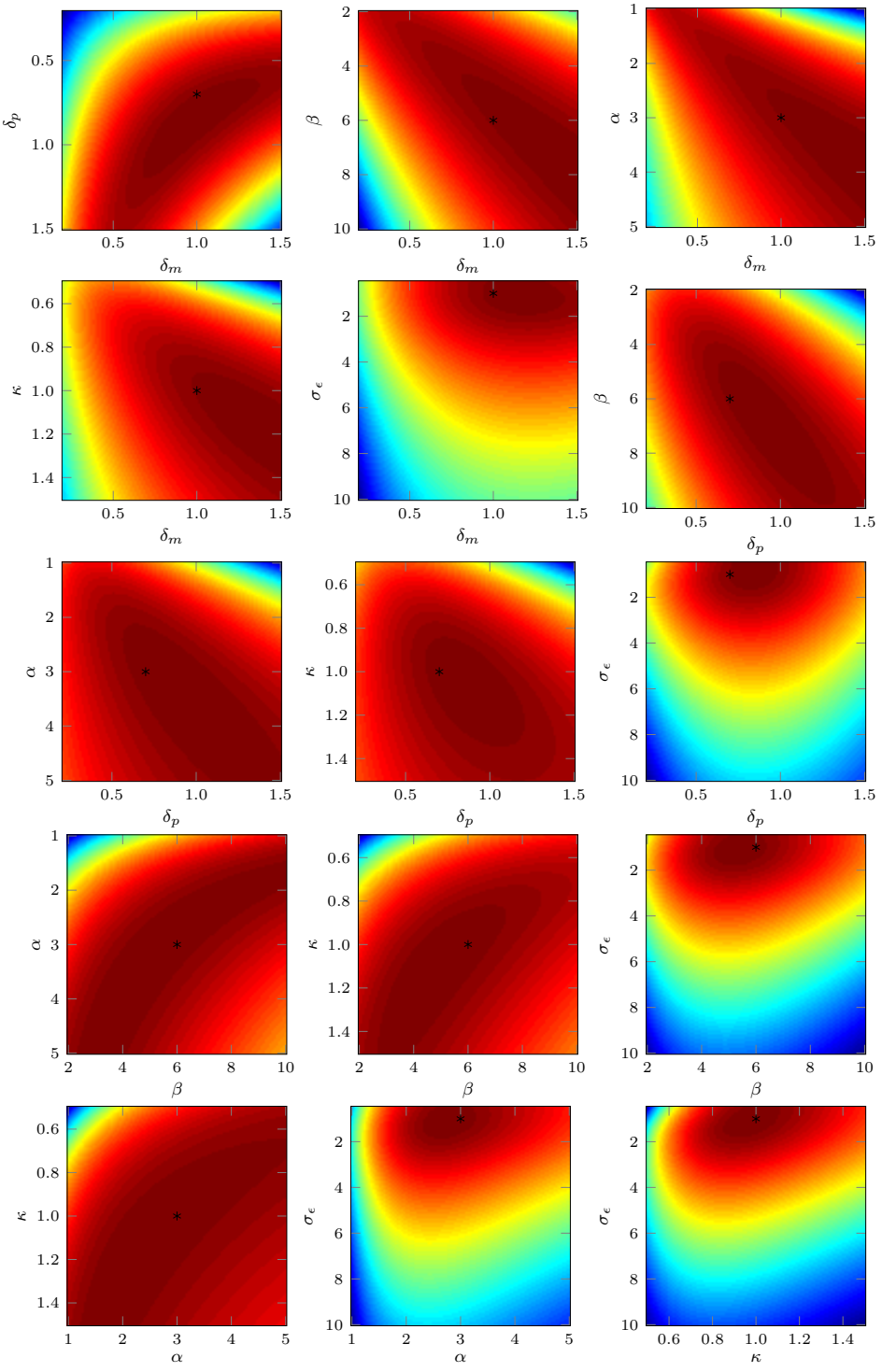


Figure 3.7: Bivariate likelihood transects under the non-restarting LNA data likelihood. True parameter values are shown by the black points. Data were generated via an exact stochastic simulation algorithm and sampled at 100 discrete timepoints.

3.5 Parameter Inference

The previous section described how to compute the data likelihood within a state space model for stochastic reaction networks under the LNA and BDA. Given this likelihood, we now focus on performing inference about the parameters θ of the multi-switch gene transcription model.

In particular, inference about θ includes inference on the number, k, and position, $s_1, ..., s_k$ of switches as well as the associated kinetic parameters $(\beta_0, ..., \beta_k, \alpha, \delta_m, \delta_p)$, the measurement parameters, $(\kappa, \sigma_\epsilon^2)$ and the initial state of the latent molecular processes, (M_0, P_0). Note that the dimension of θ varies with the number of switches, k, and thus to sample across these differing dimensions, we employ a reversible jump (RJ) scheme, outlined below in Section 3.5.2. The remaining parameters are updated via a random walk Metropolis-Hastings (MH) step, which is described in Section 3.5.1, the full MCMC algorithm for posterior inference is given in Section 3.8.

3.5.1 Random Walk Metropolis Hastings

For a given sample of switches, the kinetic rate parameters are updated via a random-walk Metropolis Hastings (RW-MH) scheme (Metropolis et al., 1953; Hastings, 1970). Since all parameters are positive, this is performed in log-space. As demonstrated in Figures 3.5 and 3.7 in Section 3.4.4, there is a strong correlation structure between certain parameters depending upon which approximation is used. Therefore, in order to sample efficiently, we re-parameterise $\widetilde{\alpha} := \kappa\alpha$ and $\widetilde{P}_{t_i} := \kappa P_{t_i}$ under the LNA (see Appendix B.2) and target the posterior of the log-parameters $(\log\theta)$ for both the LNA and BDA, as both the proposals and priors are symmetric in log-space. Efficiency was further increased through the adaptive scheme of Haario et al. (2001). Log-parameters are sampled in two blocks where proposals are drawn from a multivariate normal distribution centred at the previous value, with covariance matrix proportional to the covariance of the Markov chains up until the current iteration. This adaptation results in an ergodic Markov chain provided the target density is bounded from above and has a bounded support.

3.5.2 Reversible Jump MCMC scheme

At each iteration of the MCMC algorithm, we employ a reversible jump (Green, 1995) step in order to update the log transcriptional profile, $\log \beta(t)$, where,

$$\beta(t) := \beta_j \quad \text{for } t \in [s_j, s_{j+1}).$$

A reversible jump method is used in order to sample across the different model dimensions, corresponding to the number, k, of switch times within the transcriptional profile. This is implemented according to a similar specification as in Jenkins et al. (2013) where at each iteration of the reversible jump step, we allow one of the following three possible moves.

a) **Propose the addition** of a switch with probability b_k.
A new switch, s^* is proposed uniformly on $[0, T]$. Suppose $s^* \in [s_j, s_{j+1})$, then new values for the transcription rates are required on this interval. In particular, the new rates are defined as a perturbation of the old rates and since we are targeting the log-parameters, this is done by first drawing u uniformly on $[0, 1]$, and setting the new rates β_j^*, β_{j+1}^* so that the following set of equations are satisfied,

$$\begin{aligned} \log \beta_{j+1}^* &= \log \beta_j + u \\ \log \beta_j^* &= \log \beta_j - u, \end{aligned}$$

where β_j was the original transcription rate over the interval $[s_j, s_{j+1})$. Since the new rates have been proposed as a transformation of the old rates, $\log \beta_j$, and the random variable u, the corresponding Jacobian is given by,

$$J := \begin{vmatrix} \frac{\partial \log \beta_{j+1}^*}{\partial \log \beta_j} & \frac{\partial \log \beta_{j+1}^*}{\partial u} \\ \frac{\partial \log \beta_j^*}{\partial \log \beta_j} & \frac{\partial \log \beta_j^*}{\partial u} \end{vmatrix} = 2.$$

b) **Propose the deletion** of a switch with probability d_k.
A switch is proposed uniformly from $\{s_1, ...s_k\}$ for deletion. Suppose s_j is the candidate for deletion, then the new transcription rate, β^*, over $[s_{j-1}, s_{j+1})$ will be chosen so that,

$$\log \beta^* = (\log \beta_{j-1} + \log \beta_j)/2.$$

Associated with this transformation is the inverse Jacobian,

$$J^{-1} := \begin{vmatrix} \frac{\partial \log \beta^*_{j+1}}{\partial \log \beta_j} & \frac{\partial \log \beta^*_{j+1}}{\partial u} \\ \frac{\partial \log \beta^*_j}{\partial \log \beta_j} & \frac{\partial \log \beta^*_j}{\partial u} \end{vmatrix}^{-1} = 1/2.$$

c) **Propose to move** a switch with probability $1 - b_k - d_k$.
A candidate switch for moving is proposed uniformly from $\{s_1, ...s_k\}$, say s_j. The new placement s^*_j is proposed uniformly on the interval $[s_{j-1}, s_{j+1})$. Since there is no transformation of the rate variables associated with this move, the Jacobian is equal to 1.

As in Green (1995), we let $b_k = c\min(1, f(k+1)/f(k))$, $d_k = c\min(1, f(k-1)/f(k))$, where c is some constant (throughout our implementation, this has been fixed at 0.4) and $f(k)$ is the prior density for k switches. The proposed transcriptional profile obtained from performing one of a), b) or c) is then accepted with probability,

$$\alpha = \min(1, \text{Likelihood Ratio} \times \text{Prior Ratio} \times \text{Proposal Ratio} \times \text{Jacobian}).$$

The prior ratio is comprised of the ratio of the priors over a) the number of switches, b) the position of switches and, c) the transcriptional rates. The proposal ratio for a) a switch addition is given by,

$$\frac{d_{k+1}T}{b_k(k+1)},$$

b) a switch deletion is given by,

$$\frac{b_{k-1}k}{d_kT},$$

and c) a switch adjustment is equal to 1.

3.5.3 Hierarchical Model

Due to the strong dependency between parameters it can sometimes be difficult to disentangle them and model identifiability can become an issue. One way to tackle this issue is to incorporate external information into the prior distributions of the parameters. In the example of single cell imaging data, additional experiments can be performed to obtain estimates of the degradation parameters, δ_m and δ_p to provide informative prior distributions.

A hierarchical structure can be used to aid in the identification of the remaining parameters, since a dataset will typically consist of multiple time series from the same experiment (Finkenstädt et al., 2013). Thus, one can assume that certain parameters will be similar between time series and that the variation in kinetic parameters can be modelled through this hierarchical assumption. Finkenstädt et al. (2013) motivate this approach and in particular demonstrate how a hierarchical approach can be used to model the extrinsic noise as defined by Elowitz et al. (2002).

Let $y^{(i)}$ denote the observed time series for cell i, and $\theta^{(i)} := (\beta^{(i)}(t), \alpha^{(i)}, \delta_M^{(i)}, \delta_P^{(i)}, \kappa^{(i)}, \sigma_\epsilon^{(i)})$ denote the vector of parameters for cell i, for $i = 1, ..., N$. We assume that the translation rate, α, and measurement parameters, κ and σ_ϵ, will be similar across cells and can therefore place the following hyper-distribution over each of the parameters,

$$\begin{aligned}
\alpha^{(i)} &\sim \textit{Log-Normal}(\mu_\alpha, \sigma_\alpha^2) \\
\kappa^{(i)} &\sim \textit{Log-Normal}(\mu_\kappa, \sigma_\kappa^2) \\
\sigma_\epsilon^{(i)} &\sim \textit{Log-Normal}(\mu_\sigma, \sigma_\sigma^2), \quad \text{for} \quad i = 1, ..., N.
\end{aligned}$$

The hyper-parameters $\vartheta := (\mu_\alpha, \sigma_\alpha^2, \mu_\kappa, \sigma_\kappa^2, \mu_\sigma, \sigma_\sigma^2)$ are estimated in addition to each $\theta^{(i)}$ and due to the log-normal specification can be updated conjugately where the priors of the hyper-parameters are given by,

$$\mu_\theta | \sigma_\theta^2 \sim N\left(m, (\tfrac{\sigma_\theta}{t})^2\right), \quad \sigma_\theta^{-2} \sim \textit{Gamma}(a, b),$$

where a and b are the shape and scale parameters of the Gamma distribution.

To be explicit, consider the translation rates $\alpha := (\alpha^{(1)}, ..., \alpha^{(N)})$ for all cells $1, ..., N$, where we have assumed that,

$$\log \alpha \sim N(\mu_\alpha, \sigma_\alpha^2),$$

Thus, given observations α, the hyper-parameters, μ_α and σ_α^2, can be drawn from the following full conditional distribution,

$$\mu_\alpha | \sigma_\alpha^2, \alpha \sim N\left(m^*, (\tfrac{\sigma_\alpha}{t^*})^2\right), \quad \sigma_\alpha^{-2} | \alpha \sim \textit{Gamma}(a^*, b^*),$$

where,

$$m^* = (t^{-1}m + N\overline{\alpha}_L)/(t^{-1} + N), \quad 1/t^* = t^{-1} + N.$$
$$a^* = a + N/2, \quad b^* = b + \tfrac{1}{2}\left(Ns^2_{\alpha_L} + \left(t^{-1}N(\overline{\alpha}_L - m)^2\right)/(t^{-1} + N)\right),$$

where $\overline{\alpha}_L := \frac{1}{N}\sum \log \alpha$ and $s^2_{\alpha_L} := \frac{1}{N-1}\sum(\log\alpha - \overline{\alpha}_L)^2$. In exactly the same way, the hyper-parameters $\mu_\kappa, \sigma_\kappa, \mu_{\sigma_\epsilon}, \sigma_{\sigma_\epsilon}$ can be updated.

Specifying a hierarchical model for the transcription rates $\beta := (\beta^{(1)}, ..., \beta^{(N)})$, where $\beta^{(i)} := (\beta^{(i)}_0, ..., \beta^{(i)}_k)$ is the vector for each cell i, is less straightforward. To use the same specification as above would dilute the effect of switching events since all rates would be shrunk to a single distribution through the hierarchical specification. On the other hand, vague proper priors are not a feasible option since it gives too much prior probability to the zero switch model (Green, 1995). As an alternative, we specify a hierarchical mixture model where,

$$\log \beta^{(i)} \sim \sum_{v=1}^{V} w_{\beta_v} N(\mu_{\beta_v}, \sigma^2_{\beta_v}),$$

which reduces the hierarchical shrinkage. Without resorting to a second reversible jump, it is necessary to specify the number of components in advance. One could choose several candidates and perform model selection *a posteriori*, although we found two components sufficient to capture the variability in the data, which is supported by the biological hypothesis that transcription will typically occur at either a high or low rate. Simulations showed that if the rates truly come from a single component then this is elicited from a two component specification with one weight estimated to be very low.

The hyper parameters of the mixture prior ($w_{\beta_v}, \mu_{\beta_v}, \sigma^2_{\beta_v}$ for $v = 1, ..., V$) can also be updated conjugately. To see this, we introduce some additional notation. Following the derivation given in McLachlan and Peel (2004), let f_v be the density corresponding to component v and let $\zeta = (\zeta_1, ..., \zeta_N)$ be the vector of indicator values such that $\zeta_i = (\zeta_{i1}, ..., \zeta_{iV})$ and,

$$\zeta_{iv} = \begin{cases} 1 & \text{if the } i\text{th observation is drawn from } f_v, \\ 0 & \text{otherwise.} \end{cases}$$

Given these indicator variables, inference on $w_\beta, \mu_\beta, \sigma^2_\beta$ can be performed by a series of Gibbs steps as the prior over the transcription rates can now be written in the

following form,

$$f(\beta, \zeta|\mathbf{w}_\beta, \mu_\beta, \sigma_\beta^2) = f(\zeta|\mathbf{w}_\beta) f(\beta|\zeta, \mu_\beta, \sigma_\beta^2)$$
$$= \prod_{i=1}^{N} \prod_{v=1}^{V} \left(w_{\beta_v} f(\beta^{(i)}|\mu_{\beta_v}, \sigma_{\beta_v}^2) \right)^{\zeta_{iv}}.$$

With conjugate hyper-priors given by,

$$\begin{aligned}
&\mathbf{w}_\beta \sim \textit{Dirichlet}(c_1, ..., c_V) \\
&\zeta|\mathbf{w}_\beta \sim \textit{Multinomial}(1, \mathbf{w}_\beta) \\
&\mu_{\beta_v} \sim N(m, t) \\
&\sigma_{\beta_v}^{-2} \sim \textit{Gamma}(a, b) \quad \text{for } v = 1, ..., V.
\end{aligned}$$

Consequently, in order to update the hyper-parameters, $\mu_\beta, \sigma_\beta^2$ and $\mathbf{w}_\beta$, conditional on the observations $\log \beta$, one can sample from each of the following full conditional distributions,

1. $\mathbf{w}_\beta | \log \beta \sim \textit{Dirichlet}(c_1^*, ..., c_V^*)$,
   where, $c_v^* = \sum_{i=1}^{N} \zeta_{iv} + c_v$ for $v = 1, ..., V$.

2. $\zeta|\mathbf{w}_\beta, \log \beta \sim \textit{Multinomial}(1, \mathbf{w}_\beta^*)$,
   where, $\mathbf{w}_\beta^* := (w_{\beta_1}^*, .., w_{\beta_V}^*)$ with $w_{\beta_v}^* \propto w_{\beta_v} f(\log \beta|\mu_{\beta_v}, \sigma_{\beta_v}^2)$.

3. For $v = 1, ..., V$ sample,

   (a) $\mu_{\beta_v}|\log \beta \sim N(m^*, t^*)$,
   where, $m^* = (t^{-1}m + N_v\overline{\beta}_v)/(t^{-1} + N_v)$, and $1/t^* = t^{-1} + N_v$, with $N_v := \sum_{i=1}^{N} \zeta_{iv}$, $\overline{\beta}_v := \frac{1}{N_v} \sum_i \log \beta \times \zeta_{iv}$.

   (b) $\sigma_{\beta_v}^{-2} \sim \textit{Gamma}(a^*, b^*)$,
   where, $a^* = a + N_v/2$, $b^* = b + \frac{1}{2}(N s_v^2 + (t^{-1} N_v (\overline{\beta}_v - m)^2)/(t^{-1} + N_v)$ and $s_v^2 := \frac{1}{N_v - 1} \sum_i (\log \beta - \overline{\beta}_v)^2 \times \zeta_{iv}$.

Finally, we specify a hierarchy over the initial latent states M_0 and P_0. Specifically, it is assumed that,

$$M_0 \sim \textit{Gamma}(\mu_{m_0}, \sigma_{m_0}^2),$$
$$P_0 \sim \textit{Gamma}(\mu_{p_0}, \sigma_{p_0}^2),$$

where here, we have parameterised the Gamma distribution by its mean and variance. In contrast to the other parameters, this is no longer conjugate and consequently, we update $\mu_{m_0}, \sigma^2_{m_0}, \mu_{p_0}$ and $\sigma^2_{p_0}$ via a Metropolis-Hastings random walk scheme.

3.6 Prior Specification

Informative Priors.

As stated previously, within a single cell imaging framework, one can obtain prior information on the two degradation parameters. These priors are parameterised by log-normal distributions,

$$\log \delta_m \sim N(\mu_{\delta_m}, \sigma^2_{\delta_m}),$$
$$\log \delta_p \sim N(\mu_{\delta_p}, \sigma^2_{\delta_p}).$$

Within simulations, the parameters $\mu_{\delta_m}, \sigma_{\delta_m}, \mu_{\delta_p}, \sigma_{\delta_p}$ were all fixed at the true value. For the application to GFP imaging data in the following chapter, these values were obtained from Finkenstädt et al. (2013), where,

$$\mu_{\delta_m} = \log(0.14), \quad \sigma_{\delta_m} = 0.06,$$
$$\mu_{\delta_p} = \log(0.57), \quad \sigma_{\delta_p} = 0.06.$$

Hierarchical Priors.

As already stated, the remaining kinetic and measurement parameters were incorporated within a hierarchy with log-normal priors,

$$\log \alpha \sim N(\mu_\alpha, \sigma^2_\alpha),$$
$$\log \kappa \sim N(\mu_\kappa, \sigma^2_\kappa),$$
$$\log \sigma_\epsilon \sim N(\mu_\sigma, \sigma^2_\sigma).$$

The hyper-parameters were assigned uninformative prior distributions where the mean, $\mu_\theta | \sigma^2_\theta$, was given a $N(0, (100\sigma_\theta)^2)$ prior, and the precision, σ_θ^{-2} , was given a $Gamma(1, 0.001)$ prior, parameterised by its shape and scale.

The hierarchy over the transcription rates was also incorporated through a log-

normal specification where,

$$\log \beta \sim \sum_{v=1}^{2} \omega_v N(\mu_{\beta_v}, \sigma^2_{\beta_v}).$$

The hyper-parameters were assigned uninformative prior distributions where the mean was given a $N(0, 100^2)$ prior, and the precision was given a $Gamma(1, 0.001)$ prior while the weights of the hierarchical mixture model have a $Dirichlet(2, 2)$ prior.

In addition, the initial values of the latent states were also incorporated into a hierarchy, with gamma specification, parameterised by the mean and variance,

$$M_0 \sim Gamma(\mu_{m_0}, \sigma^2_{m_0}),$$
$$P_0 \sim Gamma(\mu_{p_0}, \sigma^2_{p_0}).$$

The hyper-parameters were again assigned uninformative prior distributions where the mean was given a $N(0, 100^2)$ prior, and the precision was given a $Gamma(1, 0.0001)$ prior, parameterised by its shape and scale.

Switch Priors.

The prior distributions over the switch parameters were chosen to be vague. Specifically, we define a negative binomial distribution over the prior number of switch points within the data conditional on the number of switches not exceeding some $k_{\max}$.

$$k \sim NegBin(\mu_k, \sigma^2_k; k_{\max}),$$
$$s_1, ..., s_k | k \sim Unif(0, T).$$

The parameter μ_k, is the prior expected number of switches and should therefore be chosen depending on the data application. For our purposes, this has been fixed at a value of 5 and $k_{\max}$ has been fixed at 20. We have chosen a negative-binomial as opposed to a Poisson prior in order to be less informative about the prior number of switches. In particular, we fix σ^2_k to be $4\mu_k$.

Following Green (1995), conditional on k switches, the switch positions $s_1, .., s_k$ are distributed as the even-numbered order statistics from $2k + 1$ points uniformly distributed on [0 T]. As noted in Boys and Giles (2007), this prior assumption corresponds to the distribution of consecutive switches, $(s_{j+1} - s_j)/T$ following a

$Beta(2, 2k)$ distribution.

3.7 Inferring the Latent States

Recall that under the BDA, the data likelihood is analytically unavailable. Consequently, there are two approaches to inference. Either one can use a pseudo-marginal approach as discussed in Section 3.4.3, where one relies on a Monte Carlo estimate of the data likelihood. Alternatively, one can perform inference on the extended system where rather than targeting the posterior, $f(\theta|\mathbf{y})$, we instead choose to target the extended posterior, $f(\theta, \mathbf{x}|\mathbf{y})$, where $\mathbf{x}$ are the latent states. This can be done in the following Gibbs sampler where at each iteration of the outer MCMC loop, the following steps are performed,

1. Sample the parameter vector θ from $f(\theta|\mathbf{y}, \mathbf{x})$.
2. Sample the latent states, $\mathbf{x}$, from the density, $f(\mathbf{x}|\mathbf{y}, \theta)$.

Step 1 is achieved through the MCMC sampler described in Section 3.5. Step 2 can be achieved through any of the methods discussed in Section 3.4.3 within a conditional framework and we specifically focus our attention to the implementation of the sequential importance sampler (SIS) of Doucet et al. (2000).

3.7.1 Particle Gibbs

It should be noted that in step 2 of the above Gibbs sampler, one samples the latent states not only conditional on the observed data, but also on the latent path used in the previous iteration of the outer MCMC loop. Thus, in order to implement the SIS or SIR algorithms, they should be incorporated into a conditional sequential Monte Carlo framework. Technical details of how to embed the SIR within a conditional framework are given in Andrieu et al. (2010, 2009). The resulting algorithm is termed Particle Gibbs and is outlined in Algorithm 3.

Algorithm 3 Particle Gibbs SIR

Initialisation Initialise the static parameters, θ and run an SMC method (for example the SIS/SIR in Algorithm 2 to obtain a sample $x_{1:T}$ and let B denote its ancestral lineage.

Update At each iteration of the MCMC, run the following conditional SMC algorithm,

1: At time t_1, for $j \neq B_1$ sample $N_p - 1$ particles from initial distribution $X_{t_1} \sim h(x_{t_1}|x_{t_0})$ to obtain a sample $x_{t_1}^{(1)}, ..., x_{t_1}^{(B_1)}, ..., x_{t_1}^{(N_p)}$. Compute the weights and normalise, where,

$$w_{t_1}^{(j)} \propto \frac{g\left(y_{t_1}|x_{t_1}^{(j)}\right) h\left(x_{t_1}^{(j)}|x_{t_0}\right)}{q\left(x_{t_1}^{(j)}|y_{t_1}\right)}.$$

2: For $i = 2, ..., T$,

a) For $j \neq B_{t_i}$, sample $X_{t_i}^{(j)} \sim q\left(.|x_{1:t_{i-1}}^{(j)}, y_{1:t_i}\right)$ to obtain N_p paths $x_{1:t_i}^{(1)}, ...$ $..., x_{1:t_i}^{(B_{1:t_i})}, ..., x_{1:t_i}^{(N_p)}$.

b) Calculate incremental importance weights,

$$w_{t_i}^{(j)} \propto w_{t_{i-1}}^{(j)} \frac{g\left(y_{t_i}|x_{t_i}^{(j)}\right) h\left(x_{t_i}^{(j)}|x_{t_{i-1}}^{(j)}\right)}{q\left(x_{t_i}^{(j)}|x_{1:t_{i-1}}^{(j)}, y_{1:t_i}\right)}.$$

c) Normalise weights,

$$W_{t_i}^{(j)} = \frac{w_{t_i}^{(j)}}{\sum_{k=1}^{N_p} w_{t_i}^{(k)}}.$$

d) Calculate the estimated effective sample size,

$$\hat{N}_{\text{eff}} = 1/\sum_{j=1}^{N_p} \left(W_{t_i}^{(j)}\right)^2.$$

If $\hat{N}_{\text{eff}} < \hat{N}_{\text{Thres}}$, resample the paths with weights, $W_{t_i}^{(j)}$, conditional on the ancestral lineage B, and set $W_{t_i}^{(j)} = \frac{1}{N_p}$ for $j = 1, ..., N_p$ through a conditional systematic resampling algorithm (Kitagawa, 1996; Andrieu et al., 2010).

3: Sample from $1, ..., N_p$ with weights $W_T^{(1)}, ..., W_T^{(N_p)}$ to obtain a sample of the full latent path $x_{1:T}^{(B)}$ consistent with the data $\mathbf{y}$, where B is the updated ancestral lineage.

3.7.2 Particle Marginal Metropolis Hastings

In contrast to particle Gibbs, the pseudo-marginal approach would target the posterior of interest, $f(\theta|\mathbf{y})$. This can be achieved through a particle marginal Metropolis Hastings (PMMH) method where the likelihood in the acceptance ratio of the Metropolis Hastings algorithm is replaced by a Monte Carlo estimate $\hat{f}(\mathbf{y}|\theta)$ to give,

$$A = \frac{\hat{f}(\mathbf{y}|\theta^*)f(\theta^*)q(\theta|\theta^*)}{\hat{f}(\mathbf{y}|\theta)f(\theta)q(\theta^*|\theta)},$$

for some proposal distribution q. As discussed in Section 3.4.3, $\hat{f}(\mathbf{y}|\theta)$ can be obtained via a number of different Monte Carlo procedures. In particular, one can again use the SIS/SIR algorithm as described in Algorithm 2. Although PMMH methods have been successfully applied to various approximations of SRNs (namely the CLE in Golightly and Wilkinson (2011)), we have not considered its implementation to the BDA. The main reasons for this are that although the PMMH approach is likely to decrease the amount of autocorrelation in the Markov chains, it is well known to have "sticky" behaviour with low acceptance probability (Andrieu et al., 2010). Moreover, the computational expense of the PMMH will be higher since the particle estimate of the likelihood has to be computed every time parameters are updated (thrice per MCMC iteration). In contrast, in the Particle Gibbs approach, the latent states need only be updated intermittently, and in our current implementation, this is done every 100 MCMC iterations. Although particle degeneracy can be a major obstacle in the Particle Gibbs methods, we have found 100 particles to be sufficient to obtain a reasonable number of independent samples of the latent states, which does not induce much computational cost. Having said this, we note the PMMH may be a useful alternative to the Particle Gibbs approach implemented in the simulation study presented in Section 3.9.

3.8 Algorithm Specification

Combining Sections 3.3-3.7 together, the algorithm specification for sampling the full posterior,

$$f(\theta^{(1)}, ..., \theta^{(N)}, \vartheta|y^{(1)}, ..., y^{(N)}),$$

of both the parameters and hyper-parameters under each of the LNA and BDA is given below.

Case 1: LNA algorithm

1. Initialisation
   (a) Initialise parameters, θ.
2. Update hyper-parameters, ϑ, from the full conditional, $f(\vartheta|\theta^{(1)},...,\theta^{(N)},y^{(1)},...,y^{(N)}) = f(\vartheta|\theta^{(1)},...,\theta^{(N)})$.
   (a) Sample $\mu_\alpha, \sigma^2_\alpha, \omega_\beta, \mu_\beta, \sigma^2_\beta, \mu_\kappa, \sigma^2_\kappa, \mu_\sigma, \sigma^2_\sigma$ from their respective conjugate distributions.
   (b) Sample $\mu_{m_0}, \sigma^2_{m_0}, \mu_{p_0}, \sigma^2_{p_0}$ via a random walk MH step.
3. For cell $i = 1,...,N$, sample $\theta^{(i)}$
   (a) Update the number and position of transcriptional switches by RJ step.
   (b) Sample $\left[\log\left(\beta_0^{(i)},..,\beta_k^{(i)},\delta_m^{(i)},\delta_p^{(i)}\right), M_0^{(i)}\right]$ parameters by a random walk MH step.
   (c) Sample $\left[\log\left(\widetilde{\alpha}^{(i)},\kappa^{(i)},\sigma_\epsilon^{(i)}\right), \widetilde{P}_0^{(i)}\right]$ parameters by a random walk MH step.
4. Repeat steps 2 and 3 until convergence.

Case 2: BDA algorithm

1. Initialisation
   (a) Initialise parameters, θ.
   (b) For cell $i = 1,...,N$, initialise the latent states $M_1^{(i)},...,M_T^{(i)},P_1^{(i)},...,P_T^{(i)}$, by a sequential importance sampler.
2. Update hyper-parameters, ϑ, from the full conditional, $f(\vartheta|\theta^{(1)},...,\theta^{(N)},y^{(1)},...,y^{(N)}) = f(\vartheta|\theta^{(1)},...,\theta^{(N)})$.
   (a) Sample $\mu_\alpha, \sigma^2_\alpha, \omega_\beta, \mu_\beta, \sigma^2_\beta, \mu_\kappa, \sigma^2_\kappa, \mu_\sigma, \sigma^2_\sigma$ from their respective conjugate distributions.
   (b) Sample $\mu_{m_0}, \sigma^2_{m_0}, \mu_{p_0}, \sigma^2_{p_0}$ via a random walk MH step.
3. For cell $i = 1,...,N$, sample $\theta^{(i)}$ and the latent states,
   (a) Update the number and position of transcriptional switches by RJ step.
   (b) Sample $\left[\log\left(\beta_0^{(i)},..,\beta_k^{(i)},\delta_m^{(i)},\delta_p^{(i)}\right), M_0^{(i)}\right]$ parameters by a random walk MH step.
   (c) Sample $\left[\log\left(\alpha^{(i)},\sigma_\epsilon^{(i)}\right), P_0^{(i)}\right]$ parameters by a random walk MH step.
   (d) Update the latent states, $M_1^{(i)},...,M_T^{(i)},P_1^{(i)},...,P_T^{(i)}$, by a particle Gibbs step.
4. Repeat steps 2 and 3 until convergence.

3.9 Simulation Study

In order to validate the methods described in this chapter, we performed a number of simulation studies on synthetic data following the gene transcription model (2.3)-(2.4). Data were simulated from the exact Markov jump process via a stochastic simulation algorithm (SSA) (Gillespie, 1977) for a variety of parameter values and with differing transcriptional switch profiles. In particular, data were constructed to replicate the main features of the observed data in Figure 1.3 with observations taken at discrete time points under a linear Gaussian measurement process.

3.9.1 Study Design

We consider three scenarios of different parameter choices relating to different underlying population levels where each dataset contains 15 time series each consisting of 100 measurements over a 30 hour period. Within each scenario, time series are simulated with a variety of switching regimes and for each scenario we performed 10 different simulations and applied both the BDA and LNA models.

Scenario 1 is simulated from the parameter set: $\log \delta_m \sim N(\log(0.4), 0.02)$, $\log \delta_p \sim N(\log(0.7), 0.02)$, $\log \beta \sim N(\log(8), 0.3)$, $\log \alpha \sim N(\log(4), 0.05)$, $\log \kappa \sim N(\log(2), 0.05)$, $\log \sigma_\epsilon \sim N(\log(4), 0.2)$. This simulation scenario corresponds to average mRNA population sizes of 20 and average protein populations of approximately 115.

Scenario 2 is simulated from the parameter set: $\log \delta_m \sim N(\log(0.4), 0.02)$, $\log \delta_p \sim N(\log(0.7), 0.02)$, $\log \beta \sim 0.5 * N(\log(2), 0.2) + 0.5 * N(\log(10), 0.1)$, $\log \alpha \sim N(\log(4), 0.05)$, $\log \kappa \sim N(\log(2), 0.05)$, $\log \sigma_\epsilon \sim N(\log(4), 0.2)$. This scenario corresponds to an average mRNA population size of 15 and approximate average protein population of 85. The key difference to Scenario 1, is the bimodality of the transcription rates.

Scenario 3 is simulated from the parameter set: $\log \delta_m \sim N(\log(0.4), 0.02)$, $\log \delta_p \sim N(\log(0.7), 0.02)$, $\log \beta \sim 0.5 * N(\log(2), 0.2) + 0.5 * N(\log(4), 0.1)$, $\log \alpha \sim N(\log(1), 0.05)$, $\log \kappa \sim N(\log(4), 0.05)$, $\log \sigma_\epsilon \sim N(\log(5), 0.2)$. This scenario corresponds to the lowest population scenario with average mRNA levels at 8 and average Protein levels at approximately 10. As with Scenario 2, the transcription rates are simulated from a mixture model with two modes.

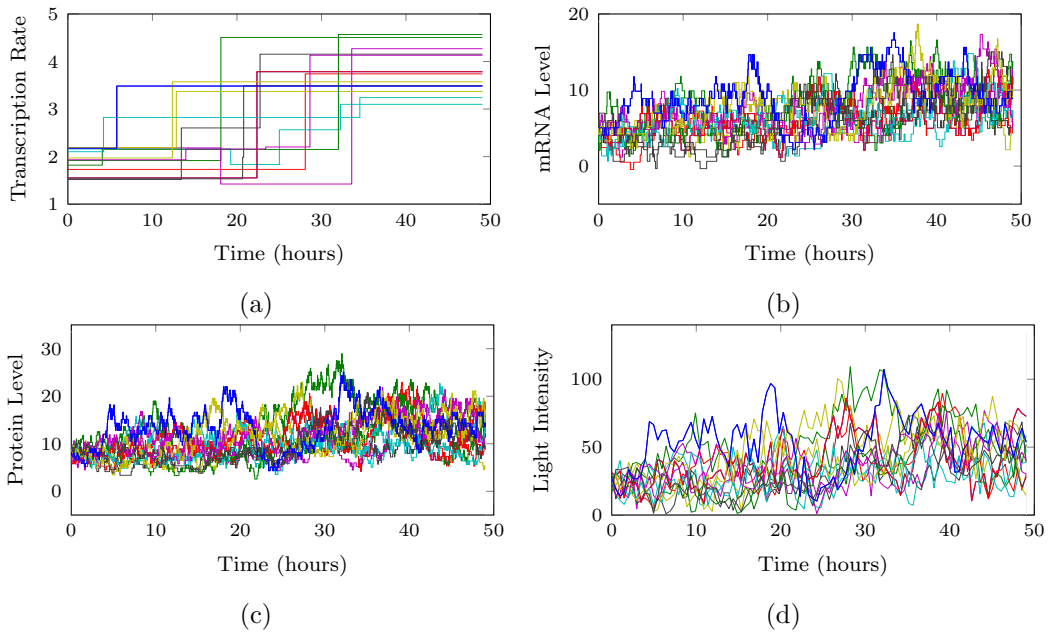


Figure 3.8: 15 simulated time series from a single hierarchical distribution under Scenario 3. a) gives the simulated transcriptional profiles, b) the corresponding continuous time mRNA process, c) the continuous time protein process and d) the discrete time observed measurements.

Applying both the LNA and BDA models to these data it was found that informative priors for the degradation parameters were essential in order to identify both the transcriptional profile, $\beta(t)$, and translation rate, α. We therefore imposed informative prior distributions, $\log \delta_m \sim N(\mu_{\delta_m}, \sigma^2_{\delta_m})$ and $\log \delta_p \sim N(\mu_{\delta_p}, \sigma^2_{\delta_p})$ where the hyper-parameters were fixed at the true values. Analyses showed that under the BDA, the scaling parameter, κ, remained unidentifiable in the majority of simulations. We hypothesise this is because under the BDA, we are targeting an extended space by explicitly sampling the latent states. To our knowledge, there has been no application within this extended framework (either explicitly or via a pseudo-marginal approach) that has incorporated a scaling parameter in the measurement equation. We hence consider two scenarios under the BDA, (1) κ is fixed at the true value and (2) κ is fixed at the posterior median obtained from the LNA.

3.9.2 Illustrative Example

We present here an example from one simulation under Scenario 3. The 15 simulated time series are shown in Figure 3.8, where a) gives the simulated transcriptional profiles, b) the unobserved mRNA levels, c) the unobserved protein levels and d)

the observed measurements. Specifically, we present the results from running the LNA and also the BDA with κ fixed at the true value.

Under the LNA, the MCMC algorithm for this simulation took 400,000 iterations compared to the BDA methodology, which took 1,500,000 iterations to sufficiently explore the respective posteriors. The corresponding thinned Markov chains (every 10 iterations) after an initial burn-in period are shown in Figure 3.9 for the LNA and Figure 3.10 for the BDA. The chains show good mixing with the true parameter value being sampled with relatively high frequency. Although the BDA was run for significantly more iterations, it can still be seen in Figure 3.10, that the parameter α has reasonably high autocorrelation in the chains. These thinned Markov chains have been used to obtain an estimate of the marginal posterior distributions, shown in Figure 3.11 (LNA) and Figure 3.12 (BDA). It can be seen that the true values all lie well within the estimated posterior densities. Moreover, the shrinking effect of the hierarchical specification can be seen, where the posterior densities for each individual parameter are very similar. This is most notable for the translation rate α, under the BDA, where the individual posterior densities are almost indistinguishable from each other. Thus, the hierarchy has a large influence on the ability to estimate the parameter α and may explain the increased autocorrelation in the Markov chains.

The true hyper-distribution for the transcription rates β, under this Scenario was a mixture of two Gaussian components. However, under both the LNA and BDA, there is convergence to a unimodal posterior for all the hyper-parameters μ_β, σ_β and also the individual parameters $\beta^{(i)}$. Specifically, it can be seen that the posterior of both the hyper means overlay each other and are unimodal where both the true modes lie well within the density. The corresponding posterior densities for $\beta^{(i)}$ are relatively wide ensuring that all the true values lie within a region of reasonable weight of the estimated posterior. This perhaps implies, that the true modes lie too close together for the algorithm to distinguish and can be seen from the data (Figure 3.8), where no distinguishable bimodality is observed in the molecular processes.

In practice, one is not only interested in the marginal posterior densities of each parameter value, but also the posterior transcriptional function, consisting of both the number and position of switch times and the corresponding transcriptional rates. Since this is estimated via a reversible jump algorithm, the posterior is a high dimensional model space consisting of all possible transcriptional functions. In order to extract the information regarding the transcriptional process, we employ a post-processing procedure outline below.

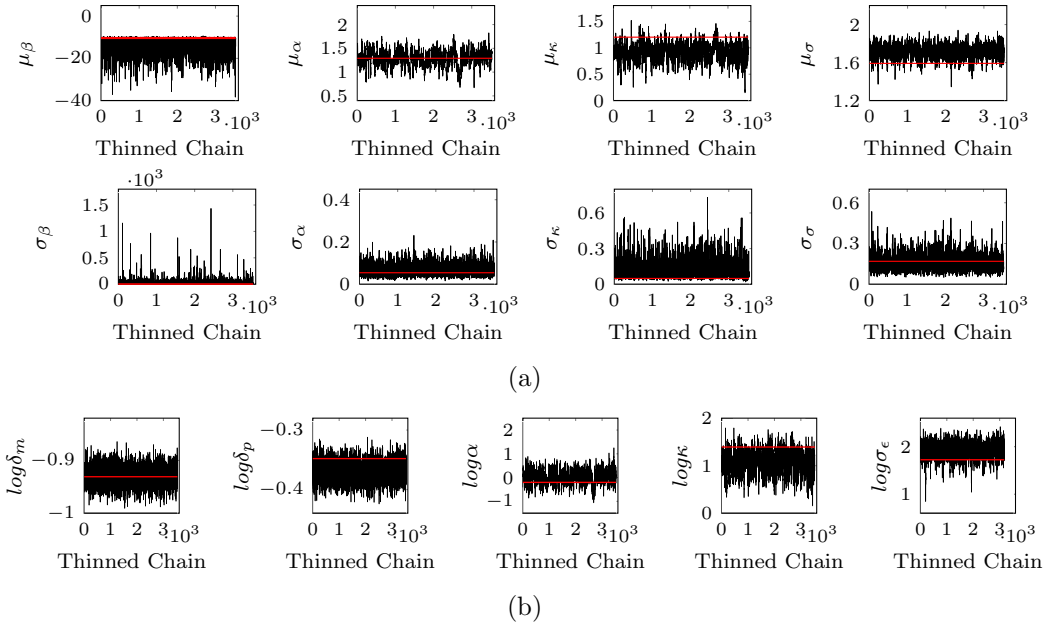


Figure 3.9: After an initial burn-in period, the thinned Markov chains for a) each of the hyper-parameters and b) each of the individual parameters for a representative single time series calculated under the LNA. Red line indicates the true value.

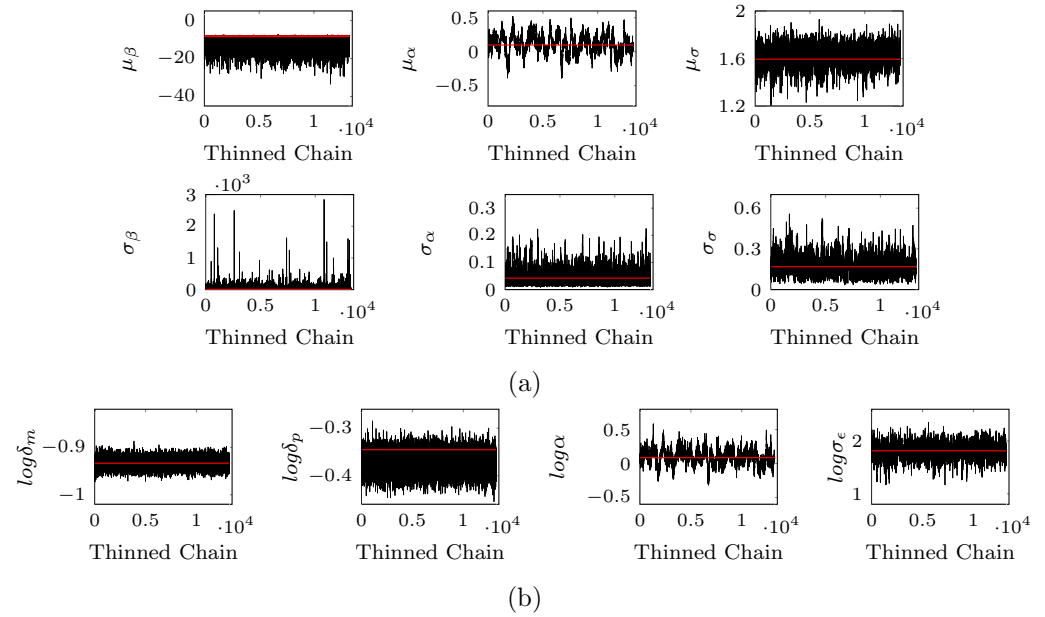


Figure 3.10: After an initial burn-in period, the thinned Markov chains for a) each of the hyper-parameters and b) each of the individual parameters for a representative single time series calculated under the BDA with red line indicating the true parameter value.

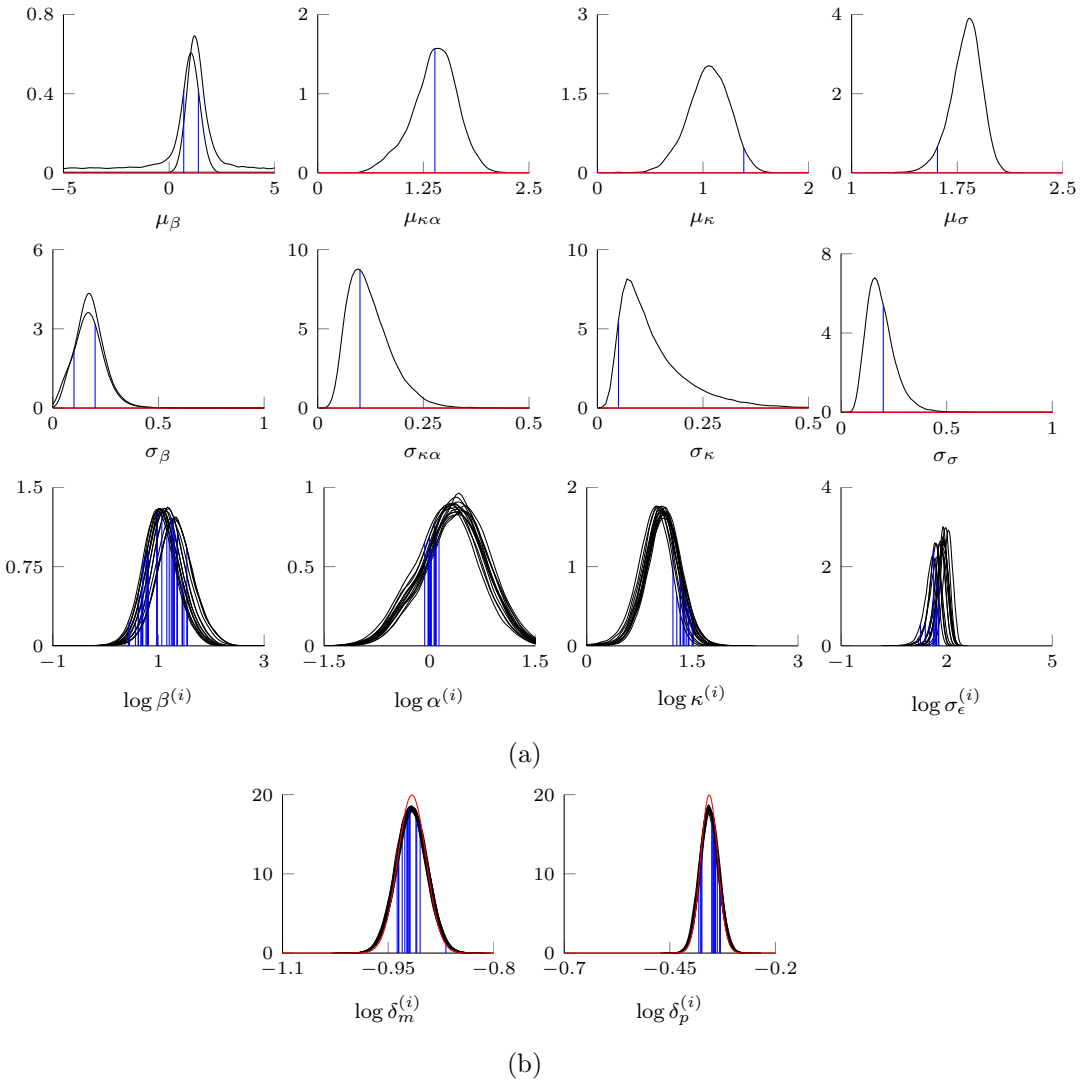
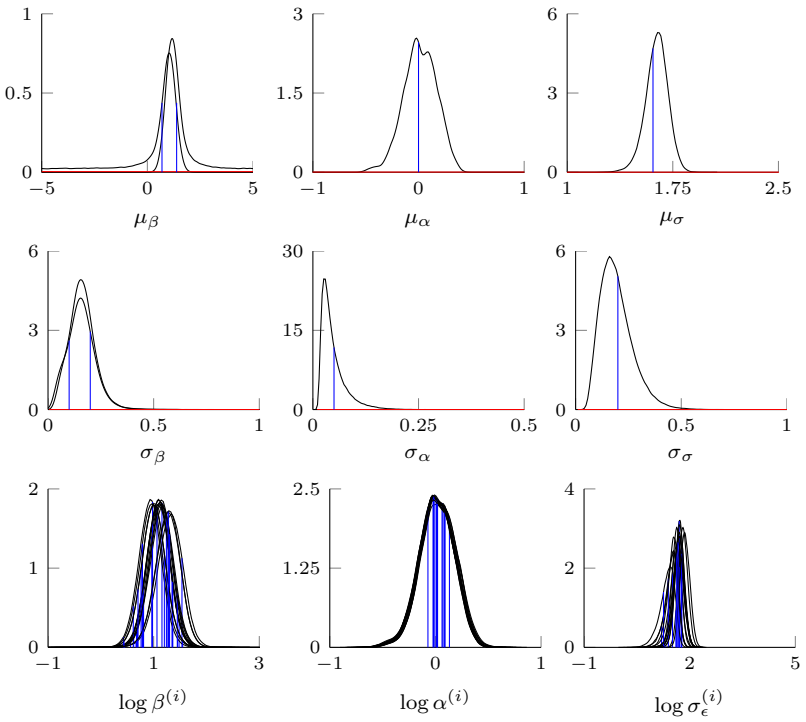
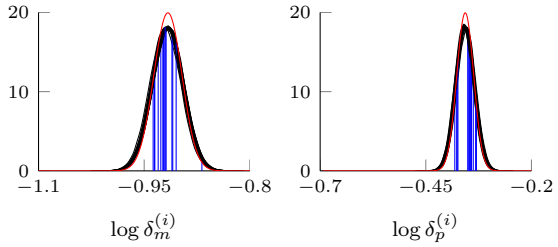


Figure 3.11: Posterior densities for a) the hierarchical parameters where the top row corresponds to the posterior density of the mean of the hyper distribution, the second row corresponds to the posterior of the standard deviation of the hyper distribution and the bottom row shows all the individual posterior densities for each parameter. In b) the posterior densities for the two degradation parameters, all of which have been calculated under the LNA. True values are shown in blue and prior densities shown in red.



(a)



(b)

Figure 3.12: Posterior densities for a) the hierarchical parameters where the top row corresponds to the posterior density of the mean of the hyper distribution, the second row corresponds to the posterior density of the standard deviation of the hyper distribution and the bottom row shows all the individual posterior densities for each parameter in the hierarchy. In b) the posterior densities for the two degradation parameters are shown, all of which calculated under the BDA. True values are shown in blue and prior densities shown in red.

Post-processing

One of the most important features to extract from these analyses is the posterior transcriptional profile. There are many ways in which this can be achieved, and for our analysis, we consider two different summaries of the output, a marginal and a conditional output. The methods for extracting these different summaries are exemplified through two different time series. The first, shown in Figure 3.13(a), has only a single weak switch in the true underlying transcriptional profile. In contrast, the second example, shown in Figure 3.14(a), has two switches (one weak, one strong) in its underlying transcriptional profile.

Marginal Transcriptional Profile

The marginal approach considers all switch points sampled during the course of the MCMC sampler, regardless of the model dimension. For example, we pool all the switch points estimated for any k switch model, $k = 0, 1, ..., k_{\max}$. These samples are shown in Figure 3.13(b) for the time series in Figure 3.13(a). In order to summarise the marginal switch points, we consider a parametric fit to the posterior density of switches in a similar vein to that presented in Jenkins et al. (2013). To do this, we first fit a non-parametric kernel density estimate, shown in Figure 3.13(c). This kernel density estimate is compared to the "null" density estimate obtained if switches were sampled completely at random, shown by the red line in Figure 3.13(c).

To test whether or not the estimated density is different from the "null", we calculate the area between the curves, shown by the shaded region in Figure 3.13(c), and if this is greater than some threshold, we conclude a non-uniform distribution of switch points. Specifically, defining,

$$\hat{f}_c(x) := \max(\hat{f}_s(x) - \hat{f}_{\text{unif}}(x), 0),$$

we conclude that the posterior switch distribution is not uniform if $\int_x \hat{f}_c(x)\ \mathrm{d}x > S_{\text{Thres}}$, where S_{Thres} is some threshold. For the purpose of this work, S_{Thres} has been set to 0.025. We note here that alternative procedures could be constructed based on the Kullback-Leibler divergence between the two densities.

If the density is different from the "null", the peaks are then extracted to give an estimate of the posterior switch times. Any peak falling below the "null" threshold is discarded, to give the new density estimate shown in Figure 3.13(d). To

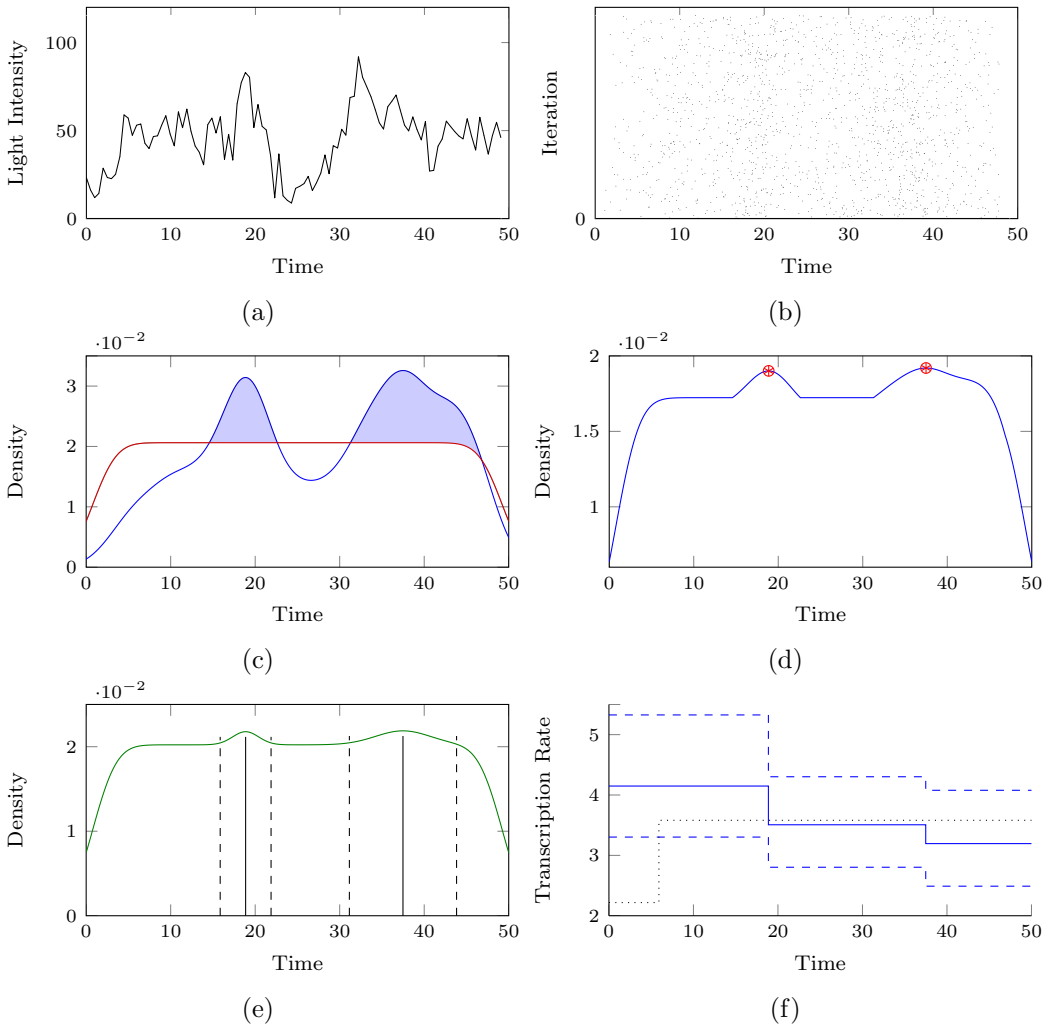


Figure 3.13: Illustrative example of fitting a marginal parametric switch model to the posterior samples obtained via a reversible jump algorithm with the LNA under Scenario 3. a) gives the raw time series data, b) shows the Markov chain for the switch position with c) a kernel density estimate of b). The red line in c) is the kernel density estimate obtained if switches were sampled uniformly over the time period. d) presents the mixture distribution of the proportion sampled uniformly and the proportion sampled in the peaks, with switch points identified by red circles. The fitted Gaussian and uniform mixture model is shown in e) with black lines indicating the estimated mean and ± 1.96 standard deviations of the switch points. The corresponding transcriptional profile is shown by the blue line in f), with the black dashed line representing the true transcriptional profile.

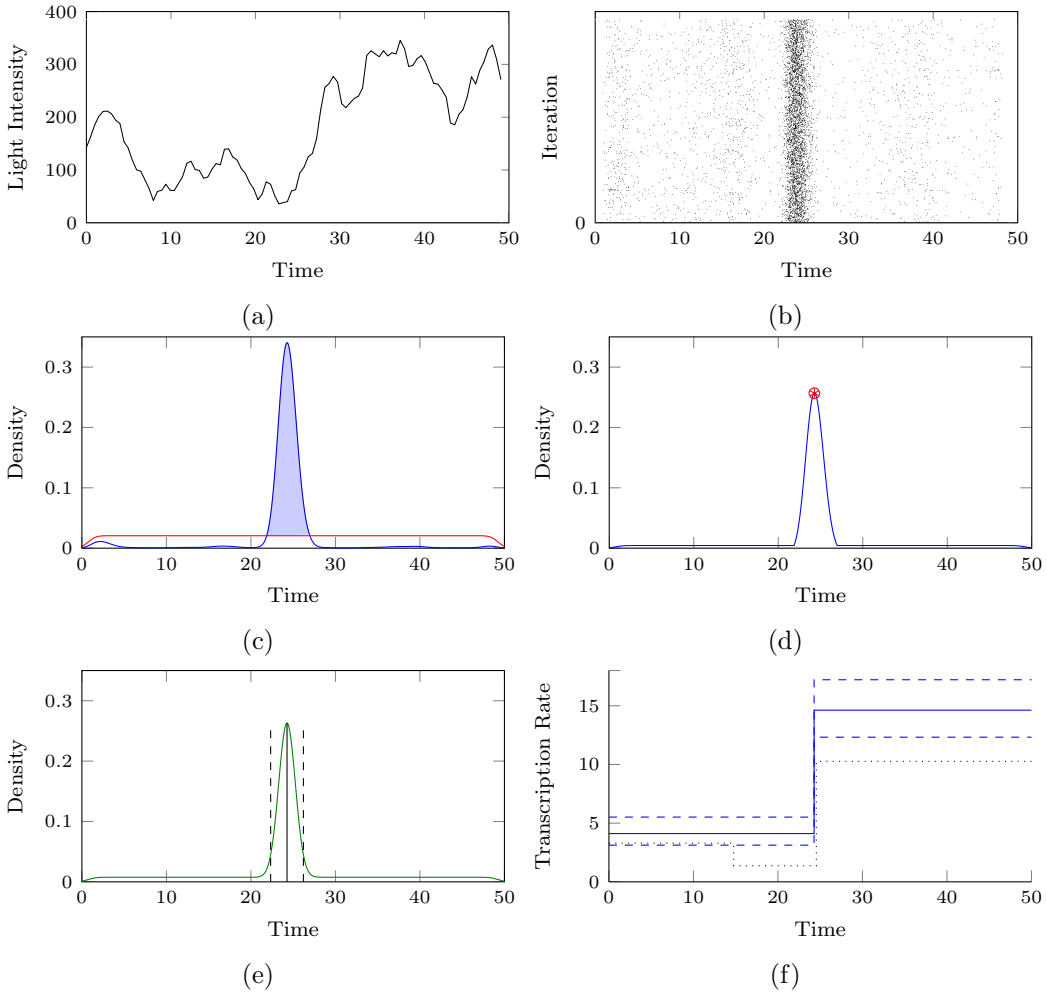


Figure 3.14: Illustrative example of fitting a marginal parametric switch model to the posterior samples obtained via a reversible jump algorithm with the LNA under Scenario 2. a) gives the raw time series data, b) shows the Markov chain for the switch position with c) a kernel density estimate of b). The red line in c) is the kernel density estimate obtained if switches were sampled uniformly over the time period. d) presents the mixture distribution of the proportion sampled uniformly and the proportion sampled in the peaks, with switch points identified by red circles. The fitted Gaussian and uniform mixture model is shown in e) with black lines indicating the estimated mean and ± 1.96 standard deviations of the switch points. The corresponding transcriptional profile is shown by the blue line in f), with the black dashed line representing the true transcriptional profile.

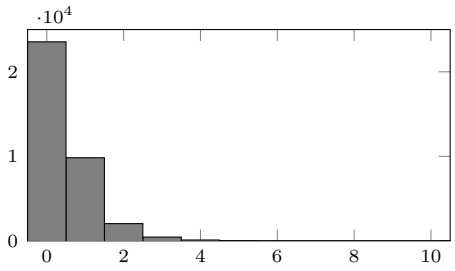


Figure 3.15: Posterior density of the number of switches sampled under the LNA for the time series shown in Figure 3.13(a).

parameterise the distribution of the switch times, a mixture model is fitted to this density. Specifically, this mixture will consist of n Gaussian components (where n is the number of peaks) and also a single component of uniform switches. The fitted mixture model is shown in Figure 3.13(e) and is fully parametric, allowing one to obtain quantities such as the standard deviation of each switch point. Associated with these marginal switch times, is a marginal transcriptional rate, which can be extracted from the reversible jump output and is shown in Figure 3.13(f).

This first example has two weak switches in the estimated posterior marginal transcriptional profile and we shall see that this summary of the algorithm output is somewhat misleading. On the other hand, running the above procedure for the second example, shown in Figure 3.14, the estimated marginal profile is found to have a single strong switch, which we will see is representative of the algorithm output.

Although, this marginal approach provides a single summary of the transcriptional profile, due to the averaging over model dimensions, substantial information is lost in this procedure. Consequently, we also consider summarising the posterior transcriptional profiles through a conditional approach.

Conditional Transcriptional Profile

The aim of a conditional summary of the posterior profile is to obtain a collection of mutually exclusive transcriptional profiles that were sampled from the reversible jump procedure along with their probabilities of occurrence. For example, the posterior distribution of the number of switches for the cell shown in Figure 3.13(a), is given in Figure 3.15 and shows that the sampler spent 65% of the time in a zero switch model, 27% of the time in a one switch model and 6% in a two switch model. Consequently, the transcriptional profile associated with the marginal two switch model rarely occurs and moreover, both the zero and the one switch models occur

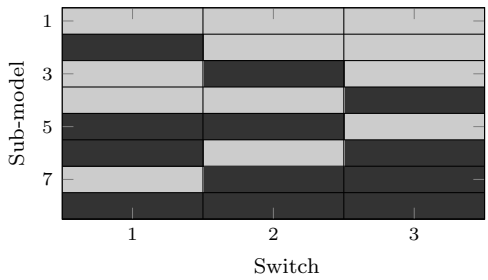


Figure 3.16: For three possible switches, there exist eight possible sub-models. Dark grey blocks represent switches that are contained in each sub-model and light grey blocks are the switches not sampled in each sub-model.

with relatively high frequency and should both be summarised.

Perhaps the most obvious method for summarising each conditional model is to first partition the posterior samples conditional on the number of switches sampled, i.e. analysing the Markov chains associated with each model dimension separately. However, this does not give a fully conditional summary and in some cases actually gives a marginal approach. Consider the extreme example where the sampler remains within a one switch dimension for the full duration of the sampler. However, suppose the single switch point is sampled in two different locations. In this example, the marginal summary will report two switches (that never co-occur) and moreover, conditioning on the one switch Markov chains, there will still be two switches estimated. Consequently, we consider a more in-depth extraction of all the conditionally mutually exclusive posterior profiles through the following procedure.

1. Extract the marginal switch distribution as outlined above, to find all possible switch points.

2. Extract the frequency of occurrence of each sub-model of the full marginal model. E.g. for a three switch marginal model, there exist eight possible sub-models as depicted in Figure 3.16. For our purposes, we perform a greedy search over this model space since the total switch dimension is typically low. Note that for more general applications, more efficient methods should be used as this model space quickly becomes very large.

3. Extract the corresponding transcriptional rates for each sub-model.

4. Report all possible models that were sampled more than $T\%$ of the time for some threshold, T.

For example, the conditional profiles for the example cell shown in Figure 3.13(a)

are given in Figure 3.17(a). In this example, the marginal model corresponds to the fourth most frequently sampled sub-model. In contrast, performing the same procedure for the example presented in Figure 3.14(a), the one switch marginal model also corresponds to the most probable model with a sampling frequency of 97.5% (shown in Figure 3.17(b)). The zero switch sub-model was only sampled 2.5% of the time.

The advantage of a conditional summary over a marginal summary, is that one retains more information, such as the number of probable models and also whether or not any of the switches are mutually exclusive. However, the conditional summary is harder to use when summarising complete datasets and also when making comparisons between different datasets. One approach we use in the following chapter, is to associate each cell with multiple profiles weighted by their posterior probability of occurrence.

Alternatively, one could extract a single transcriptional profile. This single profile could be either the marginal profile, the most probable switch profile or the profile chosen by a model selection criterion. The literature provides a vast number of potential criteria for selecting the best model, with no overwhelming consensus as to which should be used. For our purposes, working within a Bayesian framework, we consider the deviance information criterion (DIC, Spiegelhalter et al. (2002)) and its variant DIC_v, (Gelman et al., 2013). We note that although one could use a Bayes factor analysis, it is arguably less appropriate for these comparisons, since it is of interest to analyse a model conditional on certain parameters rather than a fully marginal approach.

The DIC is given by,

$$\text{DIC} = -2\log\Big(f(y^{(i)}|\bar{\theta})\Big) + 2p_D, \tag{3.24}$$

where $f(y^{(i)}|\bar{\theta})$ is the likelihood evaluated at the posterior mean of the parameters, and p_D is the effective number of parameters given by,

$$p_D = 2\big(\log(f(y^{(i)}|\bar{\theta})) - \mathbb{E}(\log f(y^{(i)}|\theta))\big),$$

where, the second term is the posterior mean of the log likelihood. If the posterior mean of the parameters, θ, is far from the mode, a negative p_D can be produced and perhaps implies that the DIC is only reasonable when the posterior is well summarised by its mean (Gelman et al., 2013).

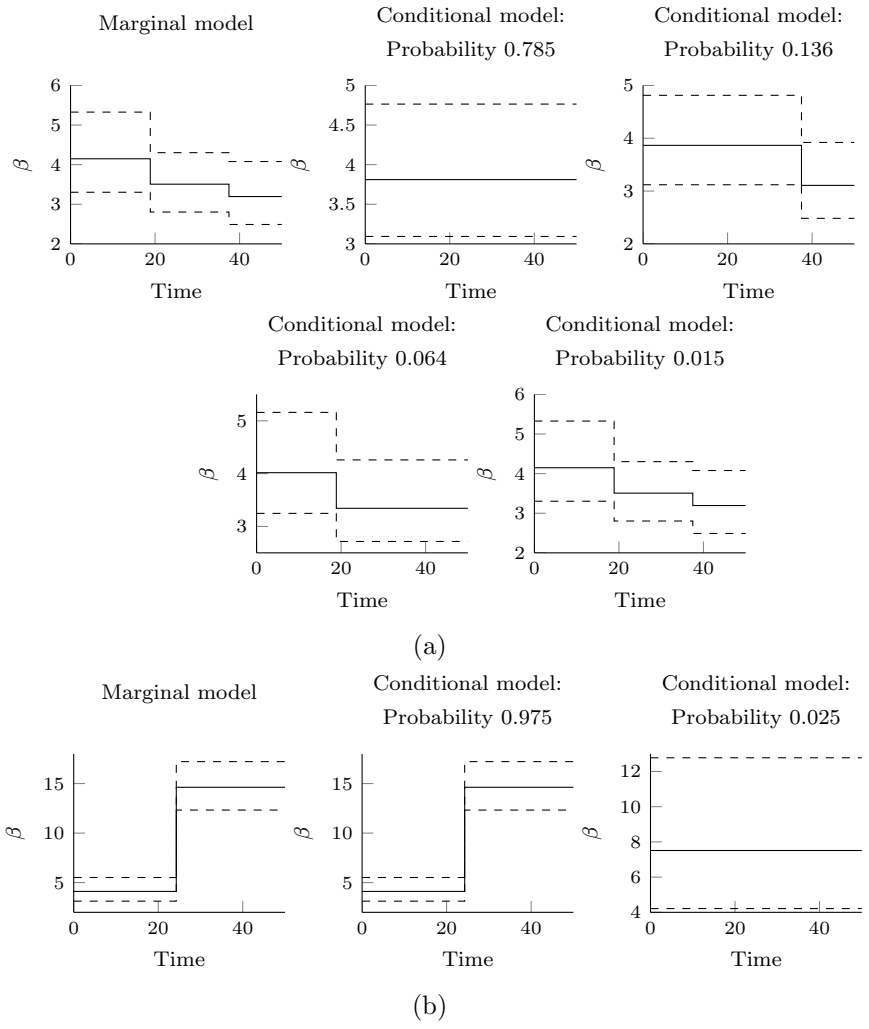


Figure 3.17: All possible sub-models for two representative cells. Panels shown in a) correspond to the single cell given in Figure 3.13 and panels in b) correspond to the single cell shown in Figure 3.14. The first panel corresponds to the marginal transcriptional profile obtained under the LNA with each following panel corresponding to an individual conditional transcriptional profile, with associated probability of occurring. Thus in the first example in a), the marginal model corresponds to the least probable model but in the second example in b), the marginal model coincides with the most probable model.

Method	Switch Model	Number of Switches	DIC	DIC_v	Proportion Sampled
LNA	Sub-Model 1	0	2.80	815.72	0.932
	Sub-Model 2	1	3.56	817.53	0.312
	Sub-Model 3	1	3.82	817.75	0.141
	Marginal Model / Sub-Model 4	2	4.31	817.60	0.0539

Table 3.1: Model Comparison of the different transcriptional profiles obtained under either a marginal or conditional extraction for the single cell presented in Figures 3.13 and Figure 3.17a). DIC is the deviance information criterion and DIC_v is its variant as given in equations (3.24) and (3.25).

As for the DIC, the DIC_v is defined as the difference between the deviance and the effective number of parameters,

$$\text{DIC}_v = -2\log(f(y^{(i)}|\bar{\theta})) + 2p_v, \tag{3.25}$$

where p_v is given by,

$$p_v = 2\,\text{Var}\left(\log(f(y^{(i)}|\theta))\right).$$

It is noted that although p_v will always be positive, it may not always be numerically stable.

These two information criteria are both derived as an approximation to the expected log pointwise predictive density, which is but one way of measuring model fit. Table 3.1 gives the DIC and DIC_v for each of the sub-models for the simulation example presented in Figure 3.17a). Unsurprisingly, the DIC is minimised for the most frequently sampled model, the zero switch model, with the marginal model having the highest DIC due to the rarity with which it was sampled.

Diagnostics

Since these inferential procedures are performed on approximations to the true underlying Markov jump process, one should assess how well these approximate models fit the data through residual analysis. For example, consider the following recursive residuals

$$r_{t_i} = \frac{y_{t_i} - \mathbb{E}(Y_{t_i}|y_{1:t_{i-1}})}{\sqrt{\text{Var}(Y_{t_i}|y_{1:t_{i-1}})}}, \quad \text{for} \quad i = 1, ..., T, \tag{3.26}$$

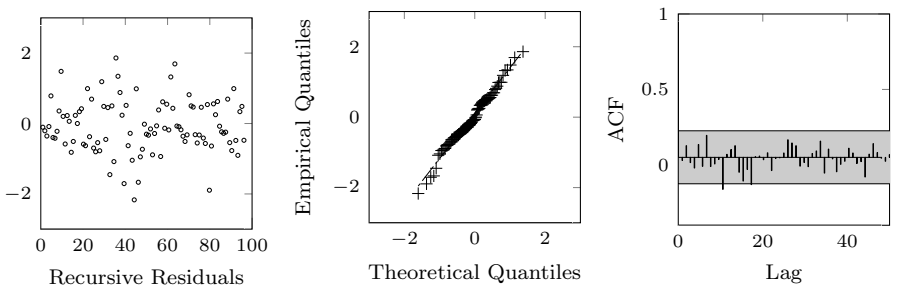


Figure 3.18: Model diagnostics from the zero switch conditional posterior model for time series shown in Figure 3.13(a) calculated under the LNA. The first panel is a plot of the recursive residuals against index, the second is a normal qq-plot and the third panel is the autocorrelation function with the grey area indicating the associated 95% confidence bands of a white noise process.

where $Y_{t_i}|y^{1:t-1}$ is the one-step ahead predictive distribution. This predictive distribution is computed in equation (3.17) as part of the Kalman filter for linear Gaussian state space models such as the LNA. It can also be evaluated via the particle filter for more general state space models such as the BDA. In particular, the moments of the predictive density can be represented by the weighted sample,

$$\mathbb{E}(h(X_{t_i})|y_{1:t_{i-1}}) = \frac{\sum_{j=1}^{N_p} w_j h(x_{t_i}^{(j)})}{\sum_{k=1}^{N_p} w_k}, \tag{3.27}$$

for weights $w_1, ..., w_{N_p}$ and samples $x_{t_i}^{(1)}, ..., x_{t_i}^{(N_p)}$ and for any function h. Under a state space model formulation the residuals in (3.26) will be independent and identically distributed with mean zero and variance one if the model fits the data. Moreover, if the state space formulation is Gaussian, these residuals will also be Gaussian.

Diagnostic plots for the chosen zero switch model of the first single cell example (Figures 3.13), are shown in Figures 3.18 and 3.19 for the LNA and BDA, respectively. The recursive residuals were computed using the posterior median of each model parameter and in both methods they appear to satisfy the state space assumptions. Namely, they have mean zero (shown in the first panel of the figures), are approximately normally distributed (qq-plot in the second panel) and are uncorrelated (autocorrelation function given in the third panel). Despite mRNA molecular numbers being less than 10, it appears from these diagnostics that the LNA is still a reasonable approach to use for inference.

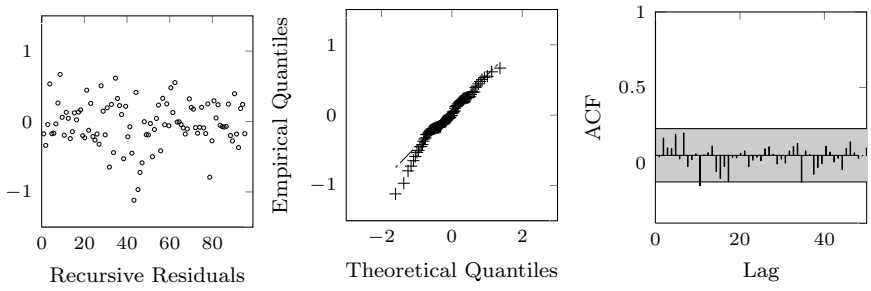


Figure 3.19: Model diagnostics from the zero switch conditional posterior model for time series shown in Figure 3.13(a) calculated under the BDA. The first panel is a plot of the recursive residuals against index, the second is a normal qq-plot and the third panel is the autocorrelation function with the grey area indicating the associated 95% confidence bands of a white noise process.

3.9.3 Results

Returning to the full simulation study, both methods were coded in MATLAB® and typically took 10-32 hours to run on a standard PC under the LNA, for 200K-700K iterations. Despite the fact that the BDA methodology is computationally faster to run per MCMC iteration, due to the high autocorrelation in the chains and poorer mixing properties, we found it would take approximately 1-3 million iterations to sufficiently explore the posterior, which could take 20-40 hours. This is unsurprising, since the BDA methodology requires the sampling of all latent states in addition to the parameter vector θ. For all scenarios under the BDA, 100 particles were used to give a sufficient number of independent samples in the particle filter.

Prior estimation of the degradation parameters is essential and moreover, the precision of these priors influences the posterior inference. Typically, 10-15 time series consisting of around 100 observations is sufficient to inform the hierarchy. More cells may be included in the hierarchy at an increased computational cost, with our methods having been successfully applied to datasets of 100 or more cells consisting of approximately 190 time points per cell (see Chapter 4).

In order to compare the two approximations, we analyse both the mean square error (MSE) and the width of the 50% credible intervals for each method. The MSE is given by,

$$\begin{aligned} \text{MSE}(\theta) &= \mathbb{E}(\hat{\theta} - \theta)^2 \\ &= \tfrac{1}{S} \sum_{s=1}^{S} (\hat{\theta}^s - \theta^s)^2, \end{aligned}$$

where $\hat{\theta}$ is an estimate of θ and $S = 150$ since we have 10×15 estimates for each parameter under each of the three different scenarios.

Figure 3.20 shows the mean square errors of the kinetic parameters calculated at the posterior median values for each of the three scenarios. We compare the following five different methods for inferring these parameters:

1. LNA,
2. LNA with κ fixed at the truth,
3. LNA with κ fixed at the LNA posterior median (obtained from method 1)),
4. BDA with κ fixed at the truth,
5. BDA with κ fixed at the LNA posterior median (obtained from method 1)).

Since the BDA cannot reliably estimate the scaling parameter, κ, it needs to be fixed *a priori*. In general, one may not know the value of κ, which motivates methods 3) and 5). A possible alternative to fixing κ would be to run the algorithm over a grid of "reasonable estimates" for κ and perform model selection.

From Figure 3.20 it can be seen that in some scenarios, the BDA provides a more accurate estimate for the transcription rate, β, and the translation rate, α, regardless of whether the LNA is calculated with κ fixed at the truth. Interestingly the LNA, although less accurate at estimating α and β, does reliably estimate the product $\alpha\beta$ and implies the LNA is less able to distinguish between these two parameters than the BDA. Moreover, Figure 3.21 shows the width of the 50% credible intervals under each of the different methods and in general, the BDA tends to give narrower intervals. The main advantage of the LNA is its computational efficiency and furthermore in practice one would be required to run the LNA to first obtain an estimate of κ before running the BDA methodology. These results show that one can use the BDA to further refine the LNA estimates of the kinetic parameters, which themselves give reasonable accuracy in reasonable computational run time.

Under both approximations, the hierarchical structure greatly aided model identifiability since it enabled the algorithm to differentiate between intrinsic variability and transcriptional switches.

In order to compare the LNA and BDA, one may wish to compare the different mathematical approximations used to compute the statistical model that is fitted to the data, for example to compare the fits of the BDA and LNA. Theoretically,

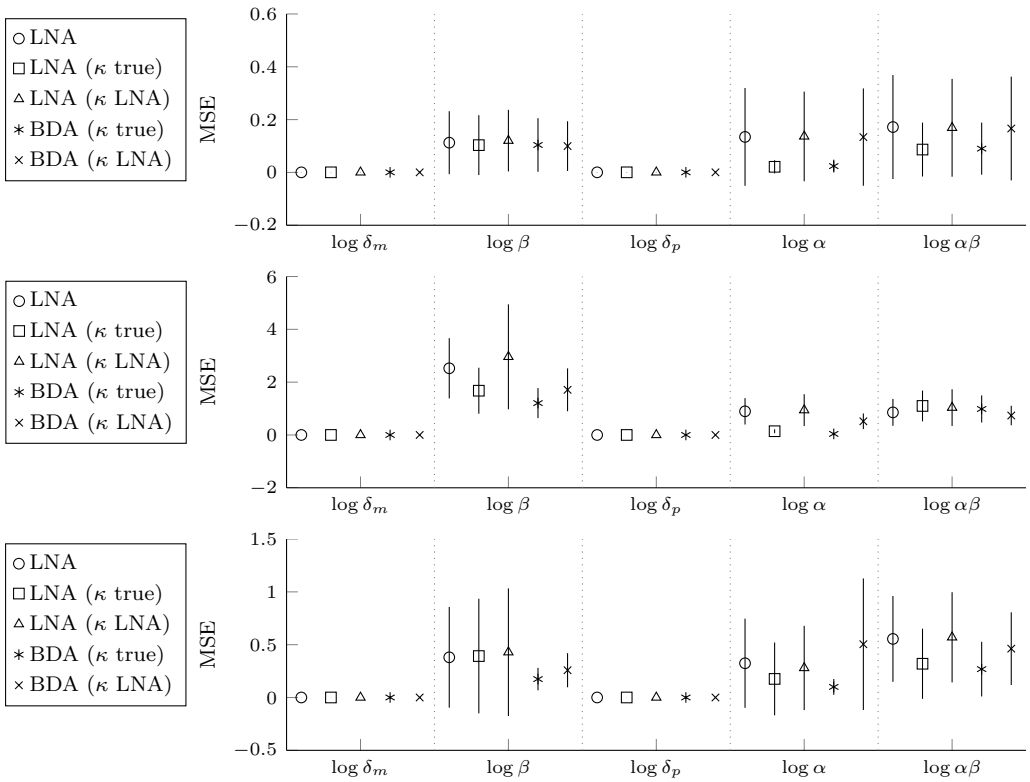


Figure 3.20: The mean square error for each estimated parameter calculated under the LNA (circle), the LNA with κ fixed at the truth (square), the LNA with κ fixed at the posterior median of the LNA (triangle), the BDA with κ fixed at the truth (star) and the BDA with κ fixed at the posterior median of the LNA (cross). Each panel corresponds to a different molecular scenario as outlined in the text. For each scenario, there are 10 different simulations containing 15 individual time series. The MSE is therefore calculated from 150 different estimates of the posterior median. The vertical lines are centred at the mean square error with length given by two standard deviations of the square error for each parameter.

one would wish to achieve this by comparing Bayes factors given by,

$$\frac{f(y|\mathcal{M}_1)}{f(y|\mathcal{M}_2)} = \frac{\int_{\theta_1} f(\theta_1|\mathcal{M}_1) f(y|\theta_1, \mathcal{M}_1) \ \mathrm{d}\theta_1 f(\mathcal{M}_1)}{\int_{\theta_2} f(\theta_2|\mathcal{M}_2) f(y|\theta_2, \mathcal{M}_2) \ \mathrm{d}\theta_2 f(\mathcal{M}_2)}, \tag{3.28}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are the candidate models with associated parameter vectors θ_1 and θ_2. Under the prior assumption that both models are equally likely, the Bayes factor reduces to the ratio of marginal model likelihoods. However, since these marginal likelihoods are not analytically available under either the LNA or BDA, one could use a Monte Carlo estimate to obtain an approximation $\hat{f}(\mathbf{y}|\mathcal{M})$. There are many different estimates suggested in the literature often obtained through an importance sampling approach. One example yields the harmonic mean estimator

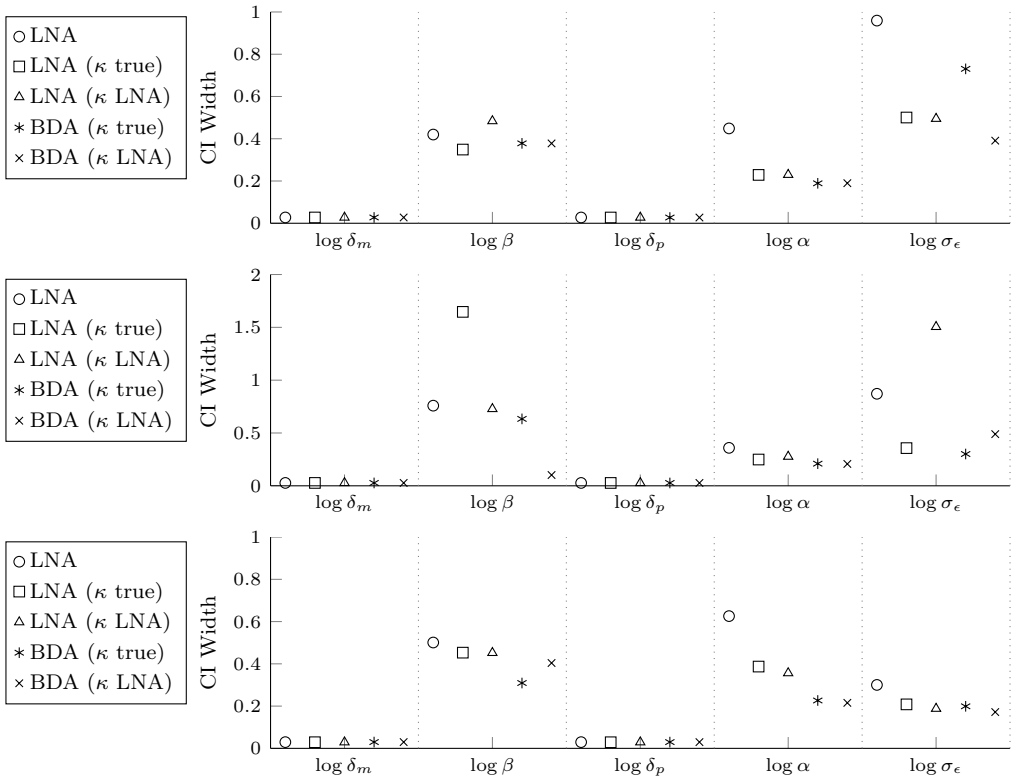


Figure 3.21: The width of the 50% credible interval calculated under the LNA (circle), the LNA with κ fixed at the truth (square), the LNA with κ fixed at the posterior median of the LNA (triangle), the BDA with κ fixed at the truth (star) and the BDA with κ fixed at the posterior median of the LNA (cross). Each panel corresponds to a different molecular scenario as outlined in the text.

(Newton and Raftery, 1994), which can be strongly affected by small likelihood values (Gamerman and Lopes, 2006). Many other importance functions have been suggested in the literature, for example see Gamerman and Lopes (2006). An alternative approach proposed by Chib and Jeliazkov (2001) is to estimate the marginal likelihood $f(y|\mathcal{M})$ through a posterior ordinate $f(\theta^*|y, \mathcal{M})$. We do not consider this a computationally feasible approach for our application, since the parameters are not updated in a single block and thus, it would require running the MCMC sampling scheme multiple times for appropriately fixed parameter values.

As an alternative to Bayes factors, model accuracy could be assessed by Bayesian cross validation. This would involve partitioning the dataset into different training and prediction sets and would again be computationally infeasible in this scenario due to the cost of performing inference. Watanabe (2010) derived an information criterion that can be viewed as an approximation to cross validation, this is given

by the WAIC,

$$\text{WAIC} = -2\left(\sum_{i=1}^{N}\log\big(\frac{1}{S}\sum_{s=1}^{S}f(y^{(i)}|\theta^s)\big) - \sum_{i=1}^{N}\text{Var}_s(\log f(y^{(i)}|\theta^s))\right),$$

where $\log f(y^{(i)}|\theta^s)$ is the log likelihood for time series i evaluated at the s iteration of the MCMC chain. Gelman et al. (2013) advocate the use of WAIC due to its fully Bayesian justification and its relation to leave one out cross validation. Note that a lower WAIC value indicates greater predictive accuracy.

In general, model comparison criteria depend upon the likelihood, which restricts its use in this modelling framework. Specifically, under the LNA, the likelihood is obtained as a marginal over the latent states whereas the BDA obtains the likelihood as a joint density over both the data and the latent states. Consequently, these model selection criteria are not directly comparable between the two approximations as can be seen by the different scales of the WAIC calculations in Figure 3.22 (comparing columns). In order to make more direct comparisons, we also obtain the joint likelihood under the LNA through the forward filtering backward-sampling algorithm described in Section 3.2.2. Thus we obtain the joint WAIC* for both the LNA and BDA methodologies defined by,

$$\text{WAIC}^* = -2\left(\sum_{i=1}^{N}\log\big(\frac{1}{S}\sum_{s=1}^{S}f(y^{(i)},x^{(i)}|\theta^s)\big) - \sum_{i=1}^{N}\text{Var}_s(\log f(y^{(i)},x^{(i)}|\theta^s))\right),$$

where $\log f(y^{(i)},x^{(i)}|\theta^s)$ is the joint log likelihood for time series i evaluated at the s iteration of the MCMC chain. These are shown in the second column of Figure 3.22 for each of the different simulations. It can be seen that in Scenario 2, the BDA methodologies generally outperform the LNA in terms of the WAIC implying it has a greater predictive accuracy. In contrast, for Scenario 3, corresponding to very low molecular numbers, the LNA surprisingly has a lower WAIC than the BDA approach. However, the numerical accuracy of the WAIC calculations can be unstable due to the low log-likelihood values. We therefore consider the diagnostic plots of Figures 3.18 and Figure 3.19 more reliable to assess the appropriateness of each approximation.

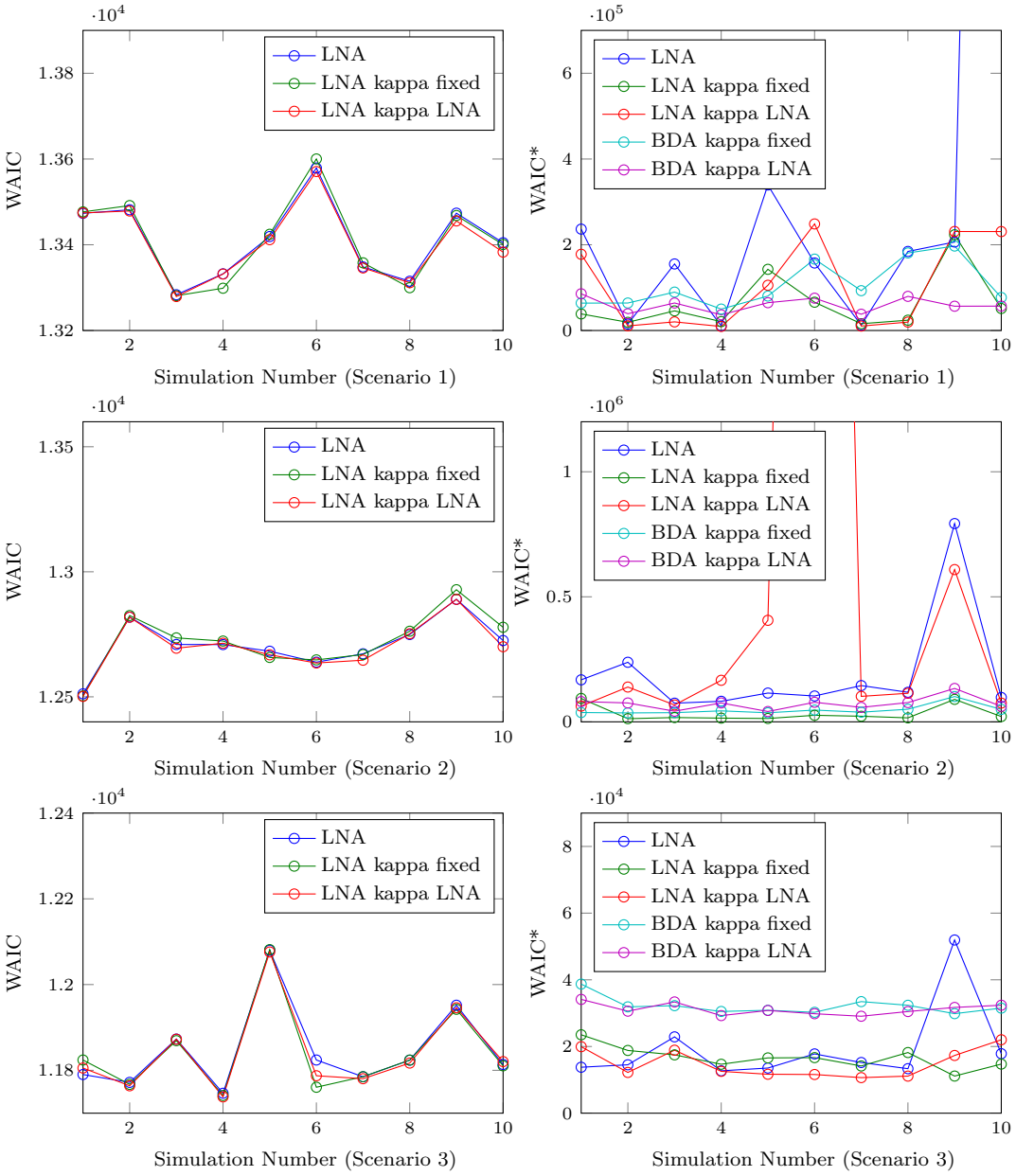


Figure 3.22: The WAIC calculated under each of the five different inference approaches. Each row corresponds to the three simulation scenarios. The first column gives the marginal WAIC of the LNA methodologies and the second column give the joint WAIC under both the the LNA and BDA methodologies. For each scenario, the WAIC is calculated for each of the 10 simulations.

3.10 Summary

In this study, we have proposed a general methodology for inferring transcriptional regulation for data obtained through single cell imaging techniques. The underlying biological model is flexible enough to describe a wide range of behaviours that cannot be captured by the traditional binary model and can be estimated reliably through a reversible jump scheme. In order to achieve the above, we consider two approximations to the true stochastic system. Although with slight loss in precision, the LNA has the advantage both in terms of computational speed, through the use of the Kalman methodology, and also its ability to identify the scaling parameter of the measurement process. This parameter is of interest as it allows one to obtain an estimate of the underlying system size. However, since the BDA can give a more accurate representation of the stochastic system, it may suggest the use of this in conjunction with the LNA estimate of κ. The BDA, although more expensive than the LNA, is still considerably cheaper than the exact methods reviewed within the previous chapter as we continue to work with the underlying transition densities albeit through a normal approximation. It therefore provides a realistic alternative to both the LNA and exact approaches when inferring systems of very small molecular numbers. The BDA is specific to our gene expression model, however, many different stochastic reaction networks can be approximated by sequences of conditionally independent birth-death reactions and a similar approach may be more widely applied. Despite the theoretical advantages of the more exact BDA, for practical implementation on large datasets we consider the LNA to give reasonable approximations in realistic computational run time.

The following chapter will discuss how these methods may be applied to real single cell imaging data and demonstrate the biological insights one can gain from such analysis.

CHAPTER 4

APPLICATION TO SINGLE CELL IMAGING DATA

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

John Tukey

The following chapter is dedicated to the analysis of several datasets measuring levels of human Prolactin expression levels within single cells across intact pituitary tissue. Data have been obtained from our collaborators at the University of Manchester and particular acknowledgement should be given to Karen Featherstone who collected all the data analysed in this chapter. Details of the specific experimental framework can be found in Semprini et al. (2009); Harper et al. (2010) and Featherstone et al. (2011).

In total, we analyse eight datasets that have been taken from mammalian pituitary tissue in different stages of development. There are two main aims of the analysis of this data, the first is to understand the transcriptional regulation of Prolactin and how this may vary in different stages of development. The second, which we will postpone to Parts II and III of this thesis, is to investigate both the spatial organisation and spatial coordination of Prolactin expressing cells in the different stages of development. The remainder of this chapter is structured as follows, Section 4.1 will briefly describe the format of the data and any necessary pre-processing procedures, Section 4.2 will present a preliminary analysis of the different datasets with Sections 4.3-4.5 presenting the analysis obtained using the stochastic switch

model methodologies as described in Chapter 3.

4.1 Pre-processing

Each dataset consists of approximately 100 single cell time series measured over a 48 hour period with measurements taken every 15 minutes. Measurements consist of the average light intensity (or average fluorescence) over the pixels within each cell. Letting $y_q^{(i)}(t)$ denote the light intensity for pixel q within cell i at time t, the data are given by,

$$\begin{aligned} y^{(i)}(t) :&= \tfrac{1}{Q^{(i)}} \sum_{q=1}^{Q^{(i)}} y_q^{(i)}(t), \\ &= \tfrac{1}{Q^{(i)}} y_{\mathrm{TOT}}^{(i)}(t), \end{aligned}$$

where $y_{\mathrm{TOT}}^{(i)}(t)$ is the total fluorescence of cell i and $Q^{(i)}$ is the number of pixels for cell i. Consequently, assuming a measurement equation of the form,

$$y^{(i)}(t) = \kappa^{(i)} p^{(i)}(t) + \epsilon^{(i)}(t), \quad \epsilon^{(i)} \sim N(0, \sigma_\epsilon^{(i)2}), \tag{4.1}$$

where $p^{(i)}(t)$ is the total number of proteins in cell i, $\epsilon^{(i)}$ is random error and $\kappa^{(i)} := \tilde{\kappa}/Q^{(i)}$, where $\tilde{\kappa}$ is the scaling parameter between the number of proteins and the light intensity of a cell. Throughout, we assume the measurement error of cell i to be independent over time and to have a Gaussian distribution.

The light intensity levels of each dataset have arbitrary units, since the microscope settings were adjusted to capture the most active dynamic range. Consequently, the scaling parameter becomes important for relating the population level of each dataset to the observed light intensity levels.

To avoid issues with light saturation, five (of the total eight) datasets were collected by recording two measurements at each time point with two differing microscope detector settings (two channels) to capture the different dynamic ranges. An example is shown in Figure 4.1, where the lower channel clearly saturates. In comparison, the higher, more sensitive channel cannot always be used as it detects substantial background autofluorescence at low light levels resulting in limited detection ability at lower intensities. For the purpose of this analysis, we combine the two channels to obtain a single time series for each cell. The combined series is shown in Figure 4.1 c). In brief, the method for combining the channels consists of:

1. Finding the range of fluorescence intensities such that there is a linear relationship between the two channels as shown in Figures 4.2 a) -b).

2. Rescaling the higher channel (channel 2) according to this linear relationship to put the two channels on the same scale (Figure 4.2 c)).

3. Combining channel 1 with the rescaled channel 2 such that below some fluorescence threshold, measurements are taken from channel 1 and above some threshold, measurements are taken from the rescaled channel 2 as depicted in Figure 4.2 d). This then creates the combined time series data shown in Figure 4.1 c).

Having two independent time series measurements for each cell enables one to assess the appropriateness of the measurement equation. In particular, under the measurement process given in equation (4.1), the relationship between the two channels should satisfy,

$$y_1 = Ay_2 + \epsilon_y, \quad \epsilon_y \sim N(0, \sigma_y^2),$$

where $A := \kappa_1/\kappa_2$ is the ratio of scaling factors and $\sigma_y^2 := \sigma_1^2 + (\kappa_1/\kappa_2)^2\sigma_2^2$. However, as can be seen from Figure 4.2, there is strong evidence of an intercept such that,

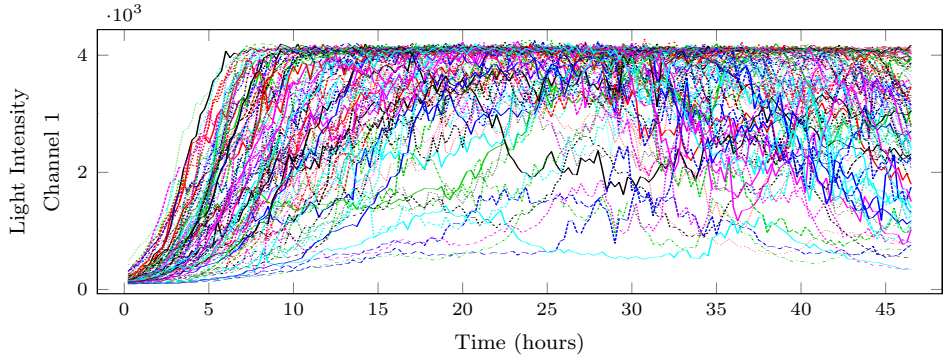
$$y_1 = I + Ay_2 + \epsilon_y, \quad \epsilon_y \sim N(0, \sigma_y^2).$$

This relationship will hold under the assumption that each individual channel measurement is also subject to an intercept term so that,

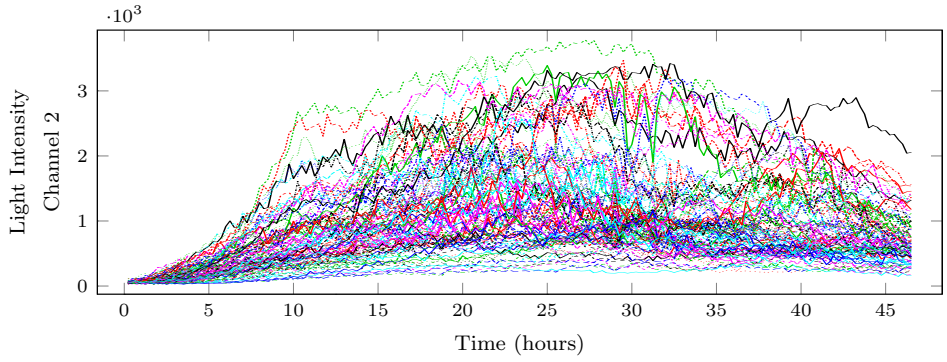
$$y(t) = \gamma + \kappa p(t) + \epsilon(t), \quad \epsilon \sim N(0, \sigma_\epsilon^2). \tag{4.2}$$

Thus, when applying to data, we include an intercept in the measurement process, which is estimated as an additional parameter. It turns out, that including this parameter in the model, greatly improves the mixing properties of the Markov chains and is well identified indicating further it is a necessary component to the measurement process. This intercept term, γ, could be the result of background autofluorescence over the tissue or a result of noise in the detector of the microscope. Thus, we impose a uniform prior over γ, limited to the range $(0, \max(\mathbf{y}))$, since the intercept will be no higher than the maximum observed light intensity measurement.

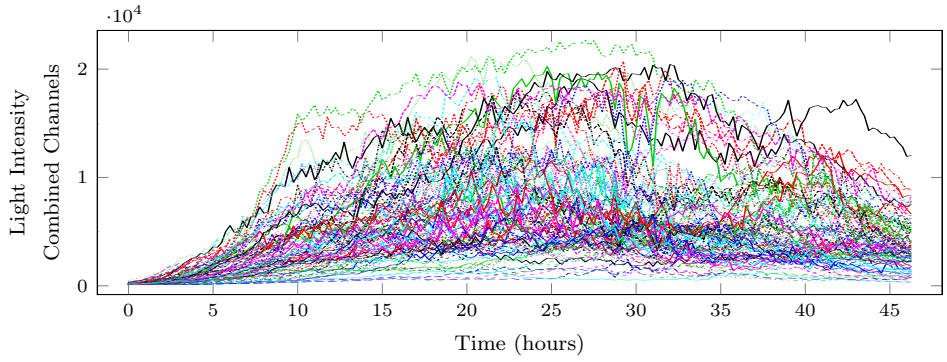
Having two channels independently measuring the same process at different fluo-



(a)



(b)



(c)

Figure 4.1: Time series data obtained from an adult male pituitary sample and imaged using two microscope detectors (two channels) to obtain two time series for each cell shown in a) and b). These have been combined to form a single time series for each cell, shown in c).

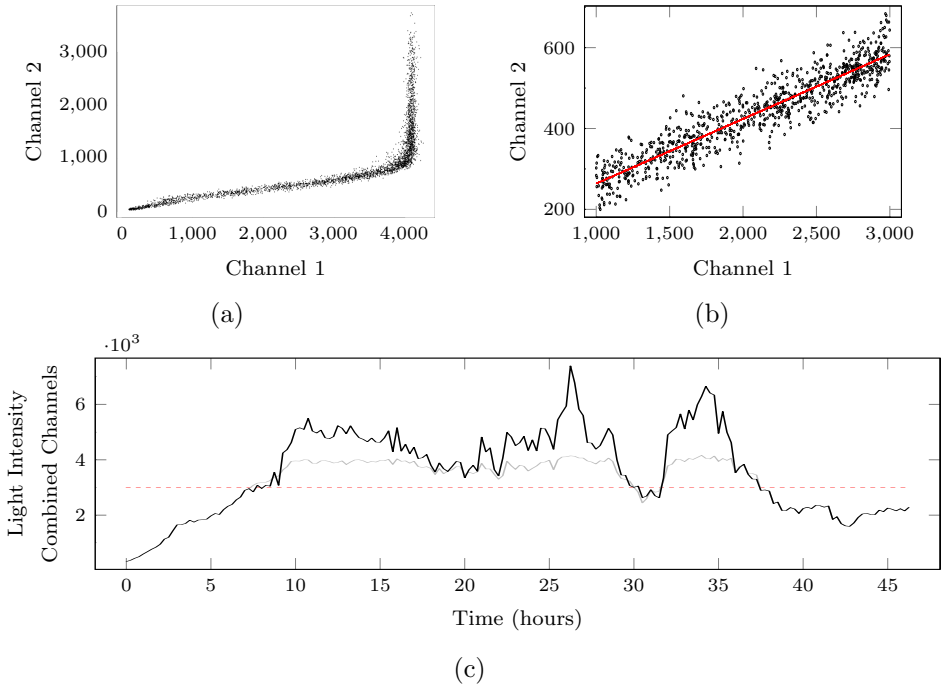


Figure 4.2: Illustration of how measurements from two different channels can be combined into a single time series. The relationship between the two channels is shown in a) with the linear relationship identified in the range shown in b). Using this relationship, the two measurements are combined in c) where the grey line indicates the channel 1 measurements and black line, the combined time series. The red line is the threshold above which the rescaled channel 2 is used.

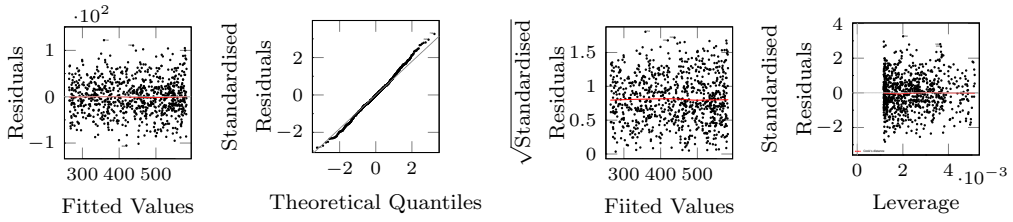


Figure 4.3: The residuals of the linear fit between the two channel measurements as shown in Figure 4.2.

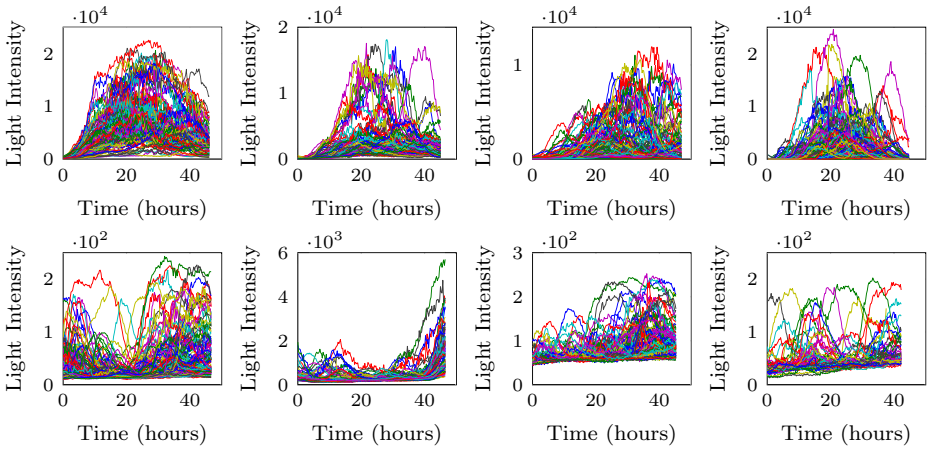


Figure 4.4: Time series imaging data for eight different datasets relating to the expression of the human Prolactin gene through a GFP reporter. Top row are four different adult male datasets consisting of approximately 100 cells each. Bottom row are two post-natal day 1.5 datasets (each of approximately 120 cells) and two embryonic day 18.5 datasets, consisting of approximately 120 and 70 cells respectively.

rescence ranges allows us to further analyse the assumptions of the measurement process. In particular, this data justifies a Gaussian error, since the residuals of the linear model in equation (4.2) are all well-behaved (Figure 4.3). This is in contrast to other findings (Finkenstädt et al., 2013), which suggest evidence of a heteroscedastic measurement process where the measurement variance is proportional to the protein population levels.

4.2 Preliminary Analysis

The eight datasets are shown in Figure 4.4 with the top row corresponding to four adult male pituitary samples and the bottom row corresponding to two post-natal day 1.5 (P1.5) and two embryonic day 18.5 (E18.5) pituitary tissues. These different datasets will be abbreviated to A1-A4 for the adult datasets, P1-P2 for the P1.5 datasets and E1-E2 for the embryonic datasets. Within each dataset, clear stochasticity can be seen due to both the measurement process and the intrinsic variability of single cell measurements. There are also clear differences in the overall trend of behaviour between the different datasets, with greater synchronicity shown in the Adult tissues. One possible hypothesis for this is a dopamine effect, which is currently being investigated through further experiments.

To get an understanding of the pulsatility and dynamics of the data, we look at both

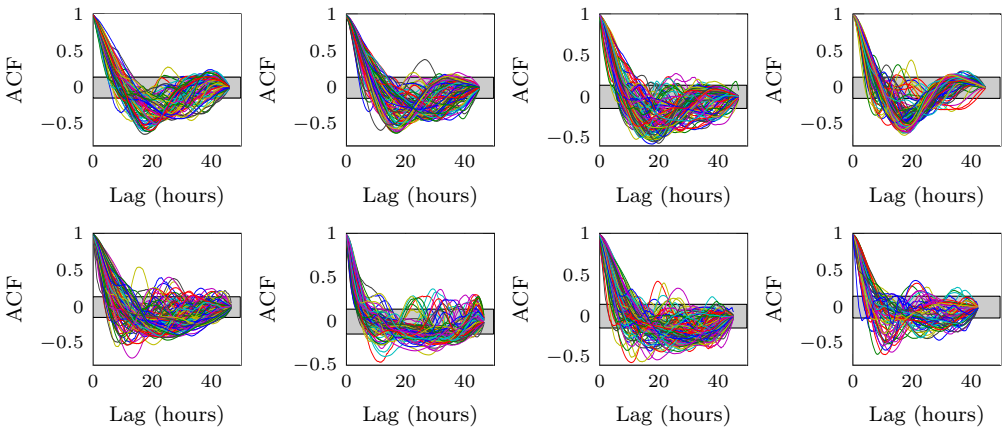


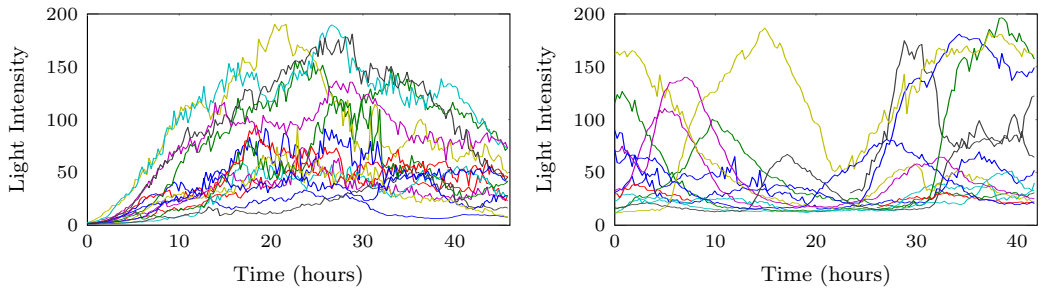
Figure 4.5: The autocorrelation function for each of the time series shown in Figure 4.4. The grey shaded region is the region within which a white noise process would be contained.

box plots of the dynamic range of expression (Figure C.1 in Appendix C) and the corresponding autocorrelation functions shown in Figure 4.5. The autocorrelation function gives clear indication of pulsatile behaviour in the sense of a deviation from white noise processes (indicated by the grey shaded region). It is difficult however to extract any meaningful phase or period due to the stochasticity of the data. In addition, the box plots are useful to show the range of activity of different cells where in most cases relatively few cells ($< 50\%$) dominate the dynamic range of the data.

These preliminary analyses allow one to make qualitative comparisons between different datasets, however, due to the arbitrary units and inability of distinguishing between intrinsic, extrinsic and measurement noise, it is difficult to compare quantitatively. In order to do this, we look at applying the stochastic switch model methodologies described in the previous chapter.

4.3 Stochastic Switch Tool Specification

In order to implement the stochastic switch tool, due to the high correlation in the likelihood surface, it was found in the simulation study that to ensure model identifiability, one requires reasonably informative priors on the two degradation rates of reporter mRNA and reporter protein. Prior information for these degradation rates was estimated in Finkenstädt et al. (2013) via two additional biological experiments.



(a) (b)

Figure 4.6: Subset of data, to illustrate the LNA and BDA methodologies. a) shows 15 randomly selected cells from dataset A1 and b) shows 15 randomly selected cells from dataset P1.

Extracting these distributions, we use the following priors,

$$\log \delta_m \sim N\left(\log(0.14), 0.06^2\right),$$
$$\log \delta_p \sim N\left(\log(0.57), 0.06^2\right).$$

We note here, that the two experiments outlined in Finkenstädt et al. (2013), to obtain estimates for reporter mRNA and reporter protein degradation rates, have also recently been performed on intact pituitary tissue with comparable degradation rates obtained.

All other prior distributions remained the same as those specified in the simulation study of the previous chapter.

4.4 Illustrative Example

To apply our methods to these data, we first consider a subset of 15 cells from datasets A1 and P1 and is shown in Figure 4.6. To back-calculate to the transcriptional dynamics, we first apply the LNA and then the BDA with κ fixed at the posterior median obtained from the LNA estimation. When applying to real data, significantly more iterations were required to fully explore the BDA posterior (8 million and 4.5 million iterations for the two datasets given in Figure 4.6 a) and b)) compared to the LNA (300K and 900K, respectively).

Figure 4.7 shows a single back-calculation under both the LNA and BDA along with the 95% credible intervals of the posterior switch times and transcription rates. This example typifies the two methods, where although the estimated transcription rates

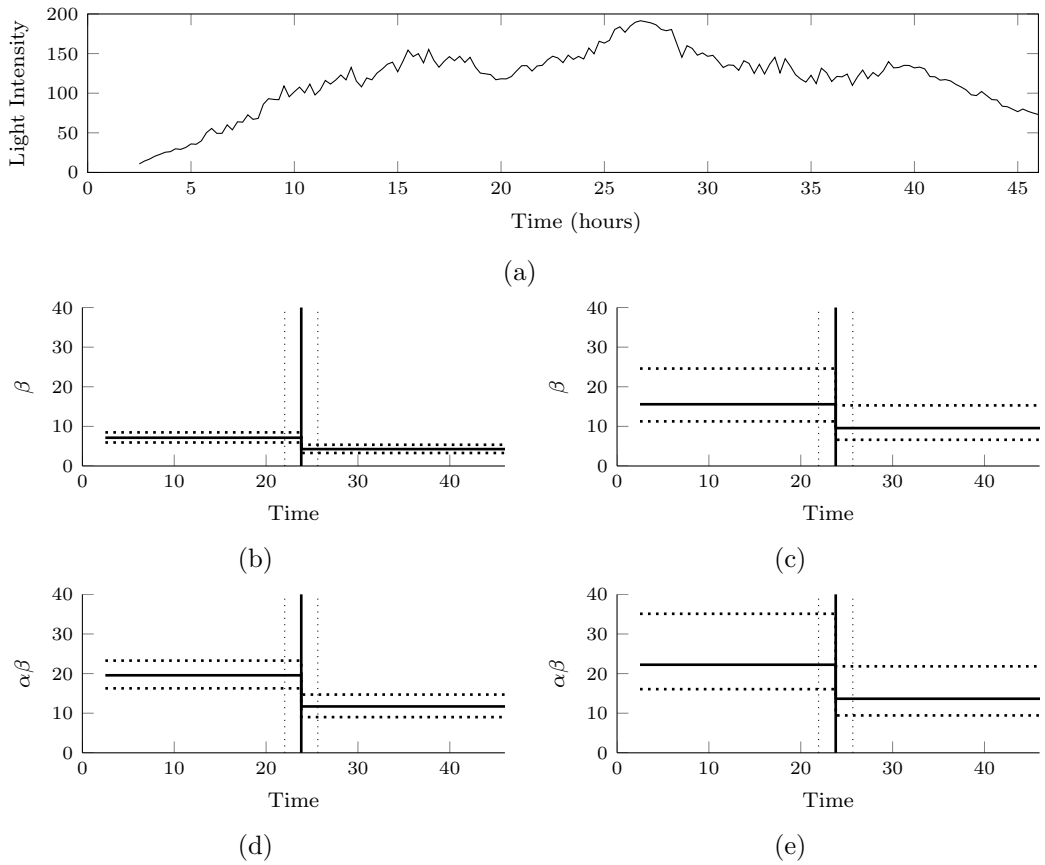


Figure 4.7: a) The raw time course data for a single cell from dataset A1, with the back-calculated transcriptional profile given in b) under the LNA and in c) under the BDA. The reparameterised profile of transcription x translation, $\alpha\beta$, is given in d) for the LNA and e) for the BDA. Dashed lines represent the 95% credible intervals about the posterior median transcriptional switches (vertical lines) and transcriptional rates (horizontal lines).

differ with tighter intervals obtained under the BDA, the product of translation and transcription, $\alpha\beta$, along with the estimated switch times are comparable. Model fit was assessed through the analysis of recursive residuals of the one-step ahead predictive distribution and are shown in Figure C.2 of Appendix C with no departure from the model assumptions detected indicating that the stochastic switch model under both the LNA and BDA fits the data well.

Also provided in Appendix C (Figures C.3-C.6) are the posterior densities obtained from running the LNA and BDA methodologies to the two datasets shown in Figure 4.6. In general, the posterior densities between the two methods show strong agreement and moreover, the importance of the prior information regarding the degradation parameters, δ_m and δ_p, can be seen. Specifically the posteriors for these parameters are indistinguishable from the informative prior densities.

Both the similarity between the LNA and BDA and the diagnostic plots suggest molecular numbers to be sufficiently large that the LNA is a reasonable approximation to the real data. One contributing factor to this, is the transgene copy number. Although unknown, for these experiments, it is thought to be in the range of 2-4 meaning for every native mRNA molecule transcribed, 2-4 reporter mRNA molecules will be transcribed resulting in larger population levels of reporter molecules.

Given the vast increase in computational expense under the BDA when applying to a subset of the real data, we present an in-depth analysis of the full data only under the LNA. Although it is possible that some identifiability between the rates of transcription and translation will be lost, given the results of the simulation study, the translation transformed transcription, given by the product $\tilde{\beta} := \beta \times \alpha$, should be well identified and comparable between datasets. Moreover, the switch times will be robust to the choice of method such that they will be comparable between datasets and representative of the true process. In order to implement the BDA in realistic computation time, this illustrative example is based on only 15 cells. In the following analysis, the estimation procedure is based on complete datasets consisting of approximately 100 cells. Consequently we see improved identifiability under the LNA since much more data is available.

4.5 Posterior Analysis

Each dataset was run through the stochastic switch model under the LNA and took approximately 300,000-600,000 iterations to sufficiently explore the posterior. Within each dataset, all (approximately 100) cells were updated in a single hierarchical structure. The marginal posterior transcriptional profiles (and translation transformed (TT) profiles) are shown in Figures 4.8 and 4.9. It can be seen that the transcriptional profiles across the different datasets exhibit dynamic switching behaviour and live on similar scales across the different datasets. It appears that the E18.5 datasets have the largest range in the estimated transcription rates, showing cells with the highest rates across all datasets. Having said this, although some cells have a very large transcription rate, there are a significant number of cells in both the E18.5 and the P1.5 datasets that remain at very low transcriptional levels throughout the time course. Dataset A1 shows the most significant switching behaviour with the majority of cells following an off-on-off type switching behaviour.

The log translation transformed marginal transcriptional profiles are shown in Figure 4.9. Plotting the data on this scale reveals evidence of a "two state" transcriptional

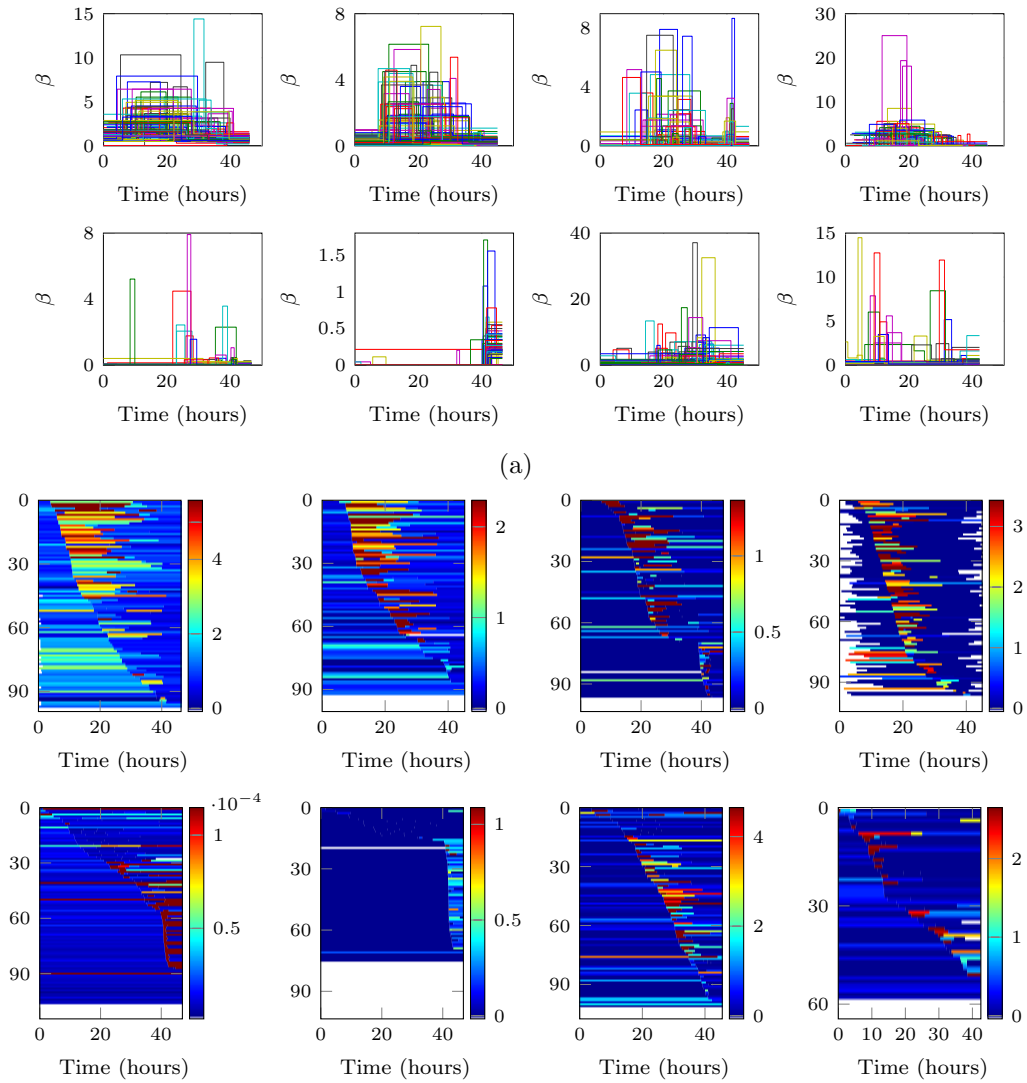


Figure 4.8: The back calculated marginal transcriptional profiles for each single cell presented as a time series in a) and a heat map in b). Each panel corresponds to the separate datasets, A1-A4 (top row), P1-P2 and E1-E2 (bottom row).

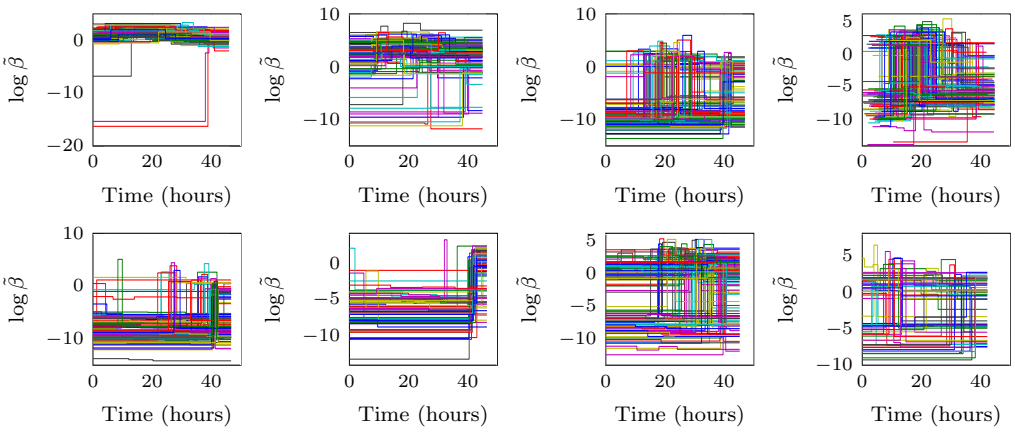


Figure 4.9: The back-calculated marginal posterior of the log translation transformed transcriptional profiles. Each panel corresponds to the separate datasets, A1-A4 (top row), P1-P2 and E1-E2 (bottom row).

model for the datasets P1-P2, E1-E2 and A2-A4, where there is a low inactive state ($\log \tilde{\beta} \in (e^{-10}, e^{-5})$) and a higher active state ($\log \tilde{\beta} \in (e^{0}, e^{6})$).

As discussed in Section 3.9, comparisons based solely on the marginal posterior profile will lose some of the information contained within the output. Specifically, the marginal switch profile may only occur with low probability where the switch times rarely co-occur. An alternative is to analyse a representative sample of transcription profiles. This can be obtained by assigning each single cell a single transcriptional profile that is randomly selected with probability given by its posterior probability of occurring. These representative samples are shown in Figure 4.10. There is a clear decrease in the dynamic behaviour with fewer switches occurring in each single cell. Moreover, in all datasets, there is a substantial increase in the number of cells that have no switches and remain at a low transcribing state throughout the entirety of the time course (Figure 4.10b)). Of those cells that switched multiple times, it is still evident that the periods of high state transcriptional behaviour in the immature (both P1.5 and E18.5) tissues is much shorter compared to the longer pulses in the Adult datasets.

In both the marginal and the representative profiles, it can be seen that there is considerable variability between datasets within the same tissue state, particularly with respect to the absolute level of transcription. For example, dataset P2, shows considerably lower transcriptional activity than P1 or in fact any other dataset. This variability makes it difficult to draw robust comparisons between the different tissue states, regarding the transcriptional level. It does appear that the two E18.5

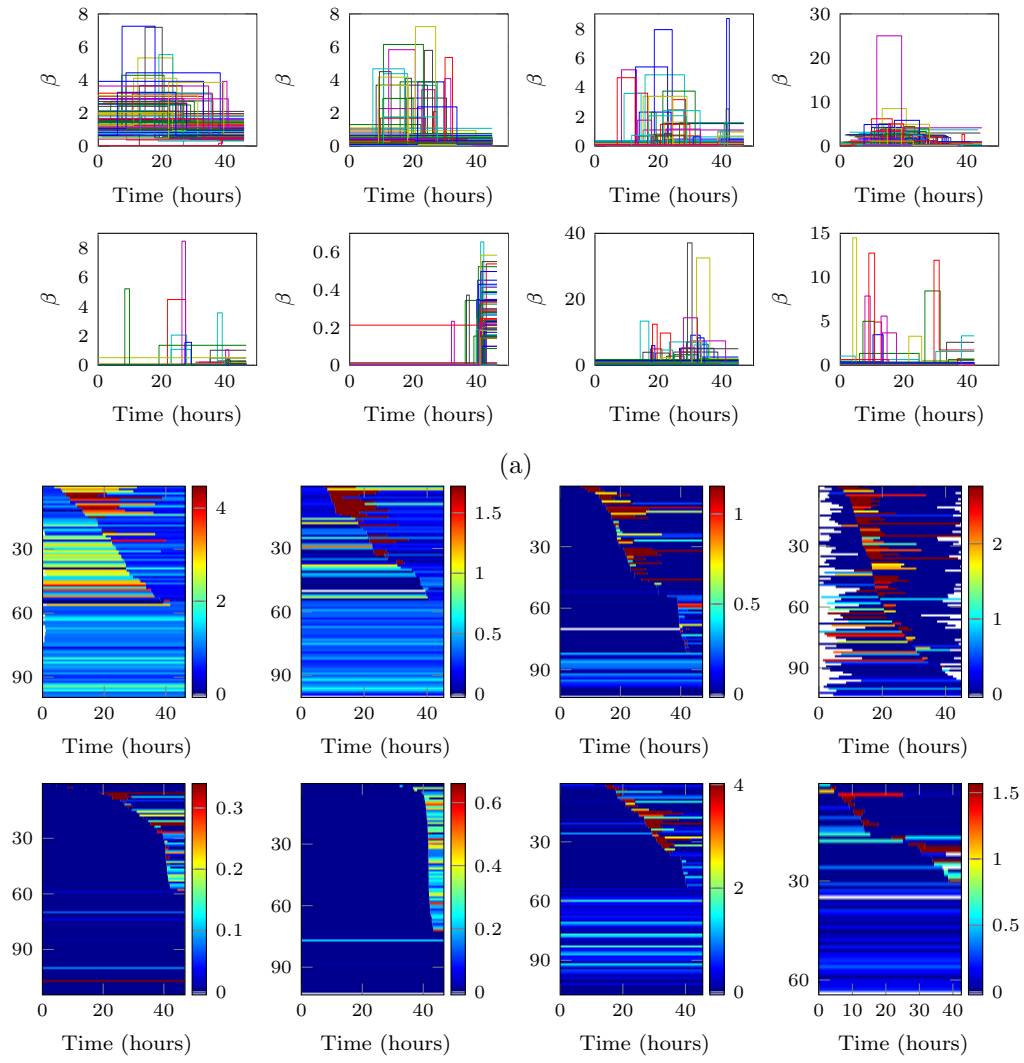


Figure 4.10: A representative sample of back calculated posterior transcriptional profiles for each single cell. The sample is obtained by randomly selecting a conditional profile with probability given by its posterior probability of occurring. These are presented as a time series in a) and a heat map in b). Each panel corresponds to the separate datasets, A1-A4 (top row), P1-P2 and E1-E2 (bottom row).

datasets, along with much shorter pulses are able to reach higher transcriptional states than the adult tissues although again, there is considerable variability between the two E18.5 datasets. Moreover, this observation is based on a relatively small sample (2 datasets) and more importantly is only based on the median transcriptional profile without taking into account the associated variability of each of the estimated profile.

Since each cell has multiple possible transcriptional profiles associated to it, using only a single (albeit representative) transcriptional profile for each cell may be insufficient to fully summarise the algorithm output. For example, some cells may have only one significant profile but others may have two or three equally likely posterior profiles. We therefore also consider summary statistics based on the weighted conditional posterior transcriptional profiles, where a summary statistic is obtained from each profile and weighted by its posterior probability of occurring. For instance, Figure 4.11 shows a) the histograms of the number of switches obtained from the marginal posterior transcriptional profiles and b) the weighted histograms of the number of switches obtained from the weighted conditional transcriptional profiles, where each cell contributes multiple values with weight given by its sampled frequency. It can be seen that the marginal approach inflates the posterior number of switches compared to the weighted conditional approach, since it does not take into account the co-occurrence of switches. In general, it seems that from the weighted conditional summary, zero, one, or two switches in transcription are sufficient to describe the pulsatility of the observed data.

In a similar way, one can analyse the posterior transcriptional rates. Figure 4.12 shows a) the histograms of the transcriptional rates from the marginal profiles and b) the weighted histograms of the transcriptional rates from all conditional posterior profiles. The shape and scale of the distributions appear to be robust to the way they are summarised with the main difference being the overall frequency. Namely, there are far fewer rates in the weighted conditional approach, due to the decrease in the number of switches. The bimodality of the log transcriptional rates is particularly interesting as it appears to capture the inactive and active ranges of transcription. Noting this is on the log scale, the lower mode essentially describes a very low basal (≈ 0) transcriptional rate and the higher mode captures a continuous range of different active transcriptional rates. This challenges the traditional binary viewpoint, since although the posterior rates can be roughly categorised into two modes (in log space), the active transcriptional rates, can take a wide continuous range of values (essentially any non-zero value) and are not restricted to a finite value.

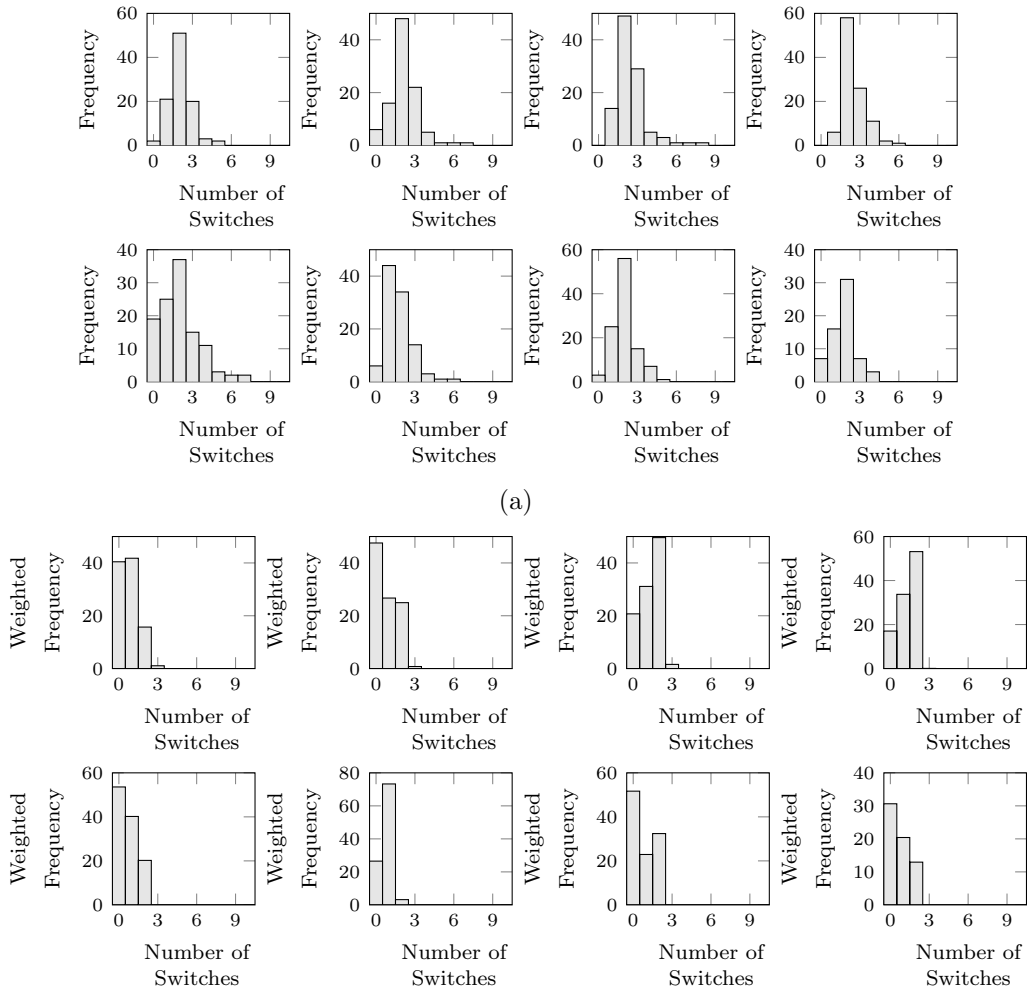


Figure 4.11: Histograms of the number of switches in the posterior transcriptional profile obtained a) from the marginal profile and b) from the weighted conditional profiles. Each panel corresponds to the datasets, A1-A4 (top row), P1-P2 and E1-E2 (bottom row).

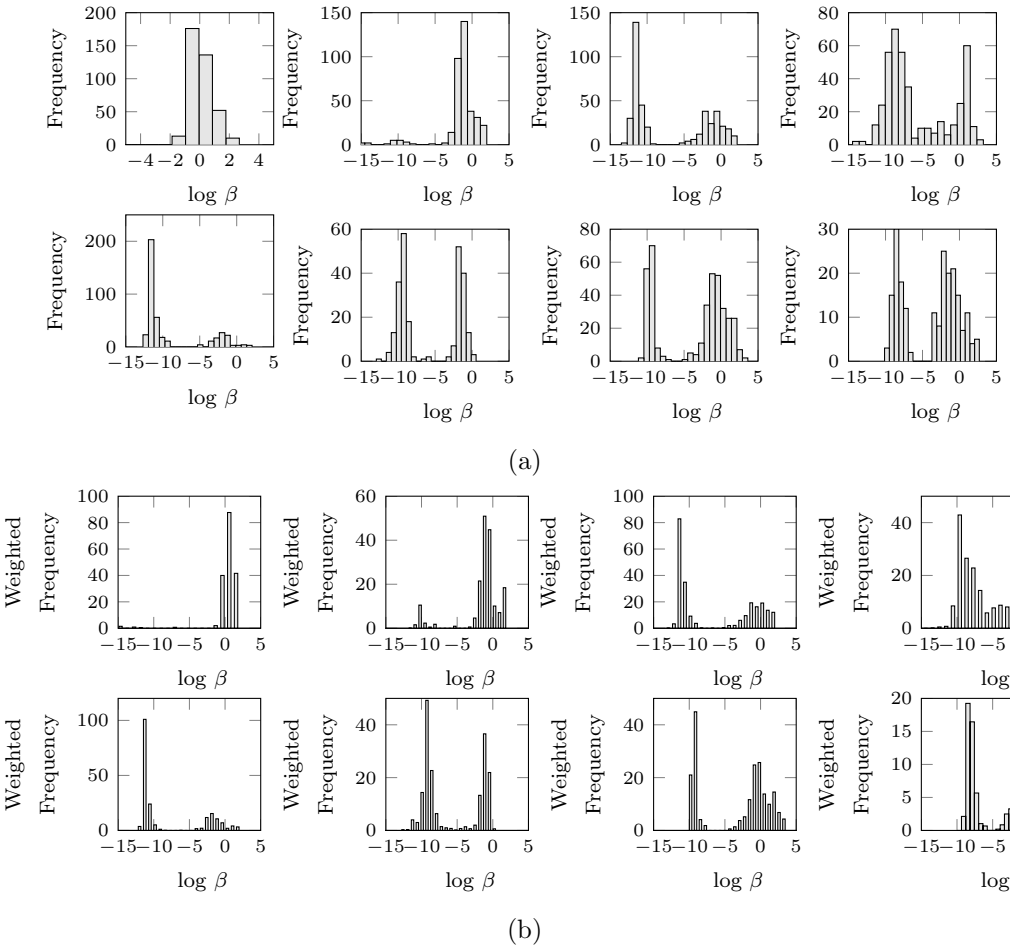


Figure 4.12: Histograms of the posterior log transcriptional rates obtained a) from the marginal profiles and b) from the weighted conditional profiles. Specifically, the histograms are obtained by counting the frequency with which a transcriptional state occurs in the posterior transcriptional profiles. Note, no weight is attached to the associated duration of each state. Each panel corresponds to the separate datasets, A1-A4 (top row), P1-P2 and E1-E2 (bottom row).

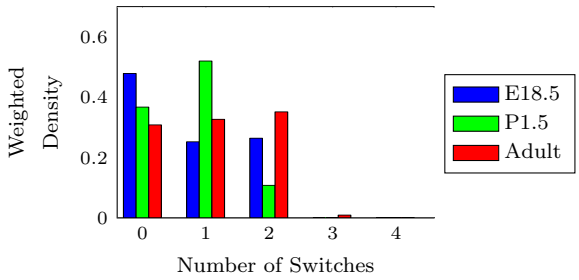


Figure 4.13: Combined histograms of the number of switches in the weighted conditional posterior transcriptional profiles for all three tissue types.

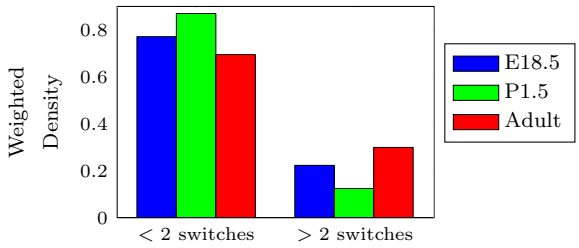


Figure 4.14: Partitions of the datasets based on whether the weighted conditional posterior transcriptional profiles have two or more transcriptional switches.

For the remainder of these analyses, we will pool together the posterior profiles into three groups of different developmental age, the four Adult datasets, the two P1.5 and the two E18.5 datasets and concentrate on comparisons between the different developmental stages. For example, Figure 4.13 shows how the number of switches differs between the different datasets and interestingly, we see significantly fewer switches in the E18.5 and P1.5 datasets compared to the Adult (Mann-Whitney U-Test with p-values < 0.001 for both E18.5 vs Adult and P1.5 vs Adult).

Partitioning the transcriptional profiles into two groups, those with fewer than two switches and those with two or more switches, one can compare the inter-switch waiting times. Figure 4.14 shows this partitioning for the different pooled datasets, from which it can be seen that there is a very small proportion of P1.5 cells that have two or more switches and consequently it may be difficult to get a reliable estimate of the waiting time between switches. Figure 4.15 shows a kernel density estimate of the inter-switch waiting times for those cells with two or more switches and it is clear from this, that both the P1.5 and E18.5 have much shorter periods compared to the adult tissues. This will be further analysed in a parametric form in the following section.

Thus, this initial analysis finds a number of interesting features within the posterior

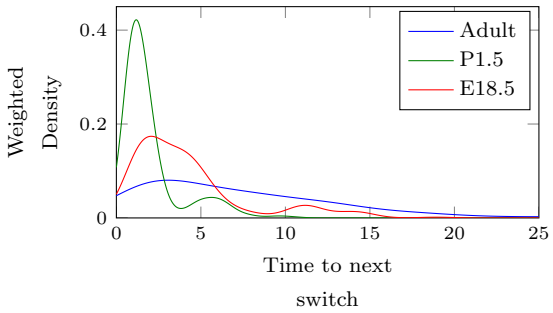


Figure 4.15: Kernel density estimate of the inter-switch waiting times for the three pooled datasets. This is calculated from all profiles with at least two transcriptional switches. Thus, the sample sizes for each pooled dataset are 150 switches for the adult, 23 switches for the P1.5 and 45 switches for the E18.5 tissues.

transcriptional profiles consisting of:

1. A continuous range of possible active transcriptional states,
2. Heterogeneous switching dynamics that differ between different development stages where:
   (a) The number of transcriptional switches in the immature tissues (E18.5 and P1.5) is lower than that in the Adult tissues,
   (b) The time between switches is shorter in the E18.5 and P1.5 tissues compared to the Adult. This is representative of the length of transcriptional pulses, since most switching cells have an off-on-off structure in their transcriptional profiles.

4.5.1 Parametric Transcriptional Process

Given these posterior transcriptional profiles and associated analysis, it is desirable to characterise the transcriptional process in a parametric form in a similar way to that already achieved under the traditional binary transcriptional model where,

$$\beta(t) = \begin{cases} \beta_{\text{on}} & \text{if a gene is active at time } t, \\ \beta_{\text{off}} & \text{if a gene is inactive at time } t. \end{cases}$$

We have already shown in the previous section that our analysis challenges this model since we find transcription can take a continuous range of different rates. However, the binary model has previously yielded parametric models for the switching dynamics. Specifically, a natural hypothesis for the binary model has been to

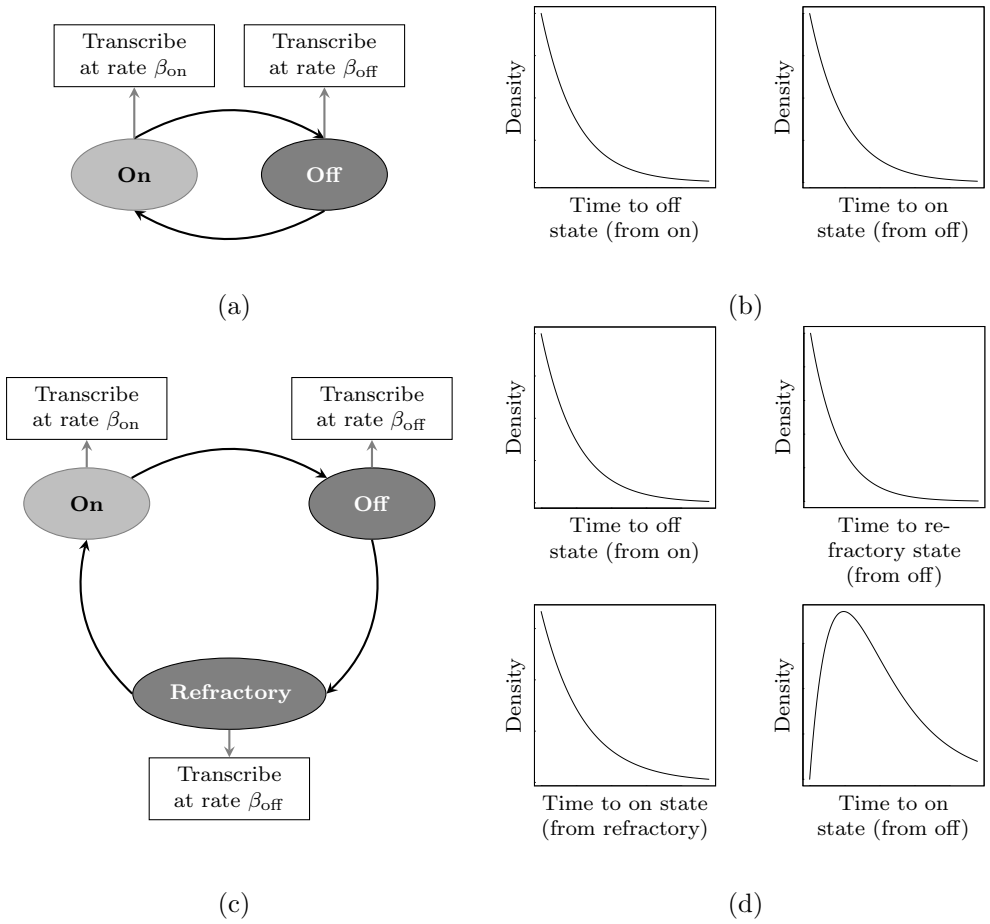


Figure 4.16: The binary switch model where a gene switches between an active and inactive state (a) with transcription occurring at a high rate in the active state and a low rate in the inactive state. The transition times between these states follow an exponential distribution (b). Incorporating a refractory state (c) where transcription remains at a low basal rate also has exponential waiting times between states (d) but the transition between an inactive (off + refractory) and an active state (on) is given by the sum of exponentials.

test if $\beta(t)$ follows a Markov process with exponential waiting times between on and off states. An interesting result from previous analyses (Suter et al., 2011; Harper et al., 2011) is evidence of a refractory period where a gene switches between three states, on-off-refractory, instead of two, on-off, and is only able to transcribe at a higher rate in the on state, as depicted in Figure 4.16. This three state model is particularly appealing as it remains Markovian with exponential waiting times between each transition but enables one to model the waiting time between an inactive state (combined both refractory and off) and an active state (on) as non-exponential. Biologically, this result translates into a cell requiring a recovery period before it is able to switch on.

Given the interpretability of the three state Markov process and its ability to capture the observed waiting time distribution, our aim is to present a similar framework to model the dynamics of observed behaviour in our more general multi-state switch model. Recall that the chosen model of the transcriptional dynamics is given by a random step function where,

$$\beta(t) = \beta_i \quad \text{for } t \in [s_{i-1}, s_i),$$

for $i = 1, ..., K$, with K being the number of transcriptional switches. Thus to parameterise the switching process, we consider two components; the time to the next switch and the rate after the next switch.

Time to Next Switch

Figure 4.17 shows the distribution of waiting times between successive switches for the pooled Adult, P1.5 and E18.5 datasets. Again we present the analysis based on both the marginal transcriptional profile and the weighted conditional profiles although the interpretation of these approaches differ substantially. Specifically, since the marginal profile can be viewed as an average transcriptional profile that combines all the mutually exclusive possibilities, the waiting times are the times between the potential switch positions and will be an underestimate of the actual waiting time between consecutive switches. In contrast, the weighted conditional approach represents the true time between consecutive switches.

There is some evidence of a departure from an exponential distribution in the waiting times of each dataset, particularly for the Adult and E18.5 pooled data (weighted marginal output). Due to the limited number of "up" switches in the back-calculated data, it is difficult to make robust comparisons between the timings of up and down switches and the distributional properties. Thus, we make no distinction in the direction of a switch in order to obtain a reasonable number of samples to enable the fitting of a parametric model.

One can capture the non-Markovian behaviour of switch times by again using the idea of a refractory period as depicted in Figure 4.18. In contrast to the binary model, which can be represented by a finite discrete state Markov process, we express the continuous switching model by an M/M/1 queuing process where a transition between two differently transcribing states consists of an exponential transition to a queue (refractory state) that is then "served" according to an additional exponential transition. Explicitly, letting R_t be the time from state β_t to the next refractory

state and S_t be the time from the refractory state to the state β_{t+1}, the full transition time, T_t will be defined as the time between the state β_t and β_{t+1} and will satisfy,

$$\begin{aligned} R_t &\sim Exp(\lambda_1), \\ S_t &\sim Exp(\lambda_2), \\ T_t &= R_t + S_t. \end{aligned} \tag{4.3}$$

Since the timings of the refractory transitions are unknown, it is impossible to decouple the two types of transitions (i.e. λ_1 and λ_2 are interchangeable), however, one can still estimate these parameters. For example, Figure 4.19 shows the likelihood surface for the waiting time data given in Figure 4.17 for a convolution of exponential distributions with parameters λ_1 and λ_2. Running a simple Newton optimisation method, we obtain the maximum likelihood estimates, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ with the corresponding fit shown by the blue lines in Figure 4.17. The estimated parameters are given in Tables 4.1 and 4.2 for the marginal transcriptional output and weighted conditional output of each pooled dataset respectively. An alternative to using the sum of exponential distributions is to approximate this by a gamma distribution with shape a and scale b^{-1}. The corresponding model fit is shown by the green lines in Figure 4.17 with parameter estimates again given in Tables 4.1 and 4.2 for the different pooled datasets. Reparameterising the MLEs in terms of the mean and variance, we see that both the Convolution model and the Gamma model estimate the same mean and comparable variances. Both appear to fit the data well, shown by the fitted curves in Figure 4.17 and moreover, there is little difference in the log-likelihood value calculated at the MLE, shown in the final column of Tables 4.1 and 4.2. These results indicate a reasonable parameterisation of the inter-switch waiting time distribution and allow direct comparisons between datasets. Under the marginal output, the Adult datasets have a mean waiting time of 6.64 hours compared to a 2 hour reduction for the P1.5 waiting times of 4.26 hours and the E18.5 waiting times of 4.93 hours, respectively. As stated previously, this will be an underestimate of the inter-switch times and can be seen when analysing the weighted conditional transcription profiles where the Adult datasets have a longer inter-switch waiting time with mean 8.82 hours. The P1.5 dataset has a mean waiting time of 3.42 hours, whilst the E18.5 remains quite robust to the marginal analysis with a mean waiting time of 4.76 hours. It should be noted, that under the weighted conditional approach, the sample size of switches for the E18.5 and P1.5 pooled data is significantly smaller than that of the Adult data at 45 and 23 switches for the E18.5 and P1.5 respectively compared to 150 samples in

Data	Fit	$\hat{\theta}$	$\hat{\mu}$	$\hat{\sigma}^2$	Log-Like
Adult	Exp C	0.16 (0.47 , -0.15) 2.15 (1.28 , 3.03)	6.64	38.31	1538.72
	Gamma	1.26 (1.69 , 0.83) 5.27 (4.87 , 5.67)	6.64	34.98	1547.92
P1.5	Exp C	0.24 (2.86 , -2.38) 22.16 (-5.06 , 49.38)	4.26	17.79	522.85
	Gamma	0.96 (1.54 , 0.39) 4.42 (3.82 , 5.03)	4.26	18.86	524.20
E18.5	Exp C	0.23 (0.64 , -0.17) 1.63 (0.77 , 2.50)	4.93	19.00	417.67
	Gamma	1.42 (1.82 , 1.02) 3.47 (3.03 , 3.92)	4.93	17.12	422.57

Table 4.1: Parameter estimates of fitting a parametric model to the waiting time distributions for the pooled Adult, P1.5 and E18.5 datasets obtained from the marginal transcriptional process. The first column gives the MLE of the parameters θ along with the 95% confidence interval. The second and third column reparameterise the parameters θ to give the mean and variance of the model respectively. The value of the log-likelihood evaluated at the MLE is given in the final column.

Data	Fit	$\hat{\theta}$	$\hat{\mu}$	$\hat{\sigma}^2$	Log-Like
Adult	Exp C	0.13 (0.40 , -0.15) 1.01 (0.56 , 1.47)	8.82	62.33	1675.39
	Gamma	1.51 (2.07 , 0.95) 5.84 (5.36 , 6.31)	8.82	51.48	1676.91
P1.5	Exp C	2.48 (0.66 , 4.29) 0.33 (0.24 , 0.42)	3.42	9.29	178.50
	Gamma	1.22 (1.76 , 0.67) 2.82 (2.13 , 3.51)	3.42	9.65	182.13
E18.5	Exp C	0.26 (0.57 , -0.06) 1.11 (0.49 , 1.73)	4.76	15.67	269.08
	Gamma	1.67 (2.00 , 1.33) 2.86 (2.37 , 3.34)	4.76	13.59	271.78

Table 4.2: Parameter estimates of fitting a parametric model to the waiting time distributions for the pooled Adult, P1.5 and E18.5 datasets obtained from the weighted conditional transcriptional process. The first column gives the MLE of the parameters θ along with the 95% confidence interval. The second and third column reparameterise the parameters θ to give the mean and variance of the model respectively. The value of the log-likelihood evaluated at the MLE is given in the final column.

the Adult. In comparison, under the marginal transcriptional approach, the sample size of the different datasets is much larger with 538, 214 and 165 switches for each of the Adult, P1.5 and E18.5 datasets respectively.

Rate after next switch

Since transcription is allowed to take any rate at any time, we investigate the relationship between consecutive rates. Figure 4.20 shows how the type of the next switch depends upon the current transcriptional rate. Unsurprisingly, the higher the current rate, the more likely it is that the next switch will be a "down" or decrease switch. In order to fully parameterise this property, Figure 4.21 constructs a density of the current transcriptional rate partitioned upon whether the next switch will be an "up" or "down" switch. Given these densities, for any given transcriptional rate, the probability that the next switch will be an increase will be given by,

$$\mathbb{P}(\text{up}|\beta) = \frac{f(\beta|\text{up})\,\mathbb{P}(\text{up})}{f(\beta)},$$

where $f(\beta|\text{up})$ is estimated by the density in Figure 4.21, $\mathbb{P}(\text{up})$ is estimated by the proportion of "up" switches and $f(\beta)$ is estimated by the sum of the densities $\hat{f}(\beta) = \hat{f}(\beta|\text{up}) + \hat{f}(\beta|\text{down})$, shown in Figure 4.21.

In general, there is a good separation between the density of the transcriptional rates conditioned upon an up switch, $f(\beta|\text{up})$, compared to the density conditioned on a down switch, $f(\beta|\text{down})$. Moreover, all the pooled datasets behave comparably. However, we again see the disadvantage to the marginal profile analysis, since there is clear bimodality in the densities that is much less evident in the weighted conditional approach. This bimodality may be artificial due to the estimation of switches that rarely co-occur in the marginal transcriptional profiles.

Given the probability of the type of the next switch, one can model the level of transcription after the next switch. Shown in Figure 4.22 is the relationship between any two consecutive rates projected to show the relationship between the lower rate versus the higher rate. One can see a weak linear relationship between the log of the lower rate and the log of the higher rate either side of a switch point such that,

$$\begin{aligned} \log \beta_{\text{low}} &= a_0 + a_1 \log \beta_{\text{high}} + \epsilon, \\ \epsilon &\sim N(0, \sigma^2). \end{aligned} \tag{4.4}$$

		Estimate	Std. Error	t value	Pr(>|t|)
Adult	a_0	-2.19	0.09	-23.78	0.00
	a_1	0.51	0.08	6.40	0.00
P1.5	a_0	-10.09	0.16	-63.37	0.00
	a_1	0.13	0.09	1.43	0.15
E18.5	a_0	-2.61	0.18	-14.90	0.00
	a_1	0.69	0.09	7.70	0.00

Table 4.3: Parameter estimates of the linear regression model (4.4) calculated from the weighted conditional transcriptional process for each pooled dataset.

This relationship is unlikely to be a complete representation of the relationship between consecutive rates, particularly due to the heteroscedasticity evident in the Adult and P1.5 datasets, however, it provides a parsimonious representation of the weak relationship seen in Figure 4.22. The corresponding parameter estimates are given in Table 4.3 and are comparable between the different datasets. The dependence between consecutive rates is most notable in the Adult and E18.5 datasets with the largest estimates of the parameter a_1. The weakest dependence is seen in the P1.5 data, which may be due to the limited number of estimated switch points. This representation will be used for simulation purposes in Chapter 7.

Rate and Switch Dependence

The final component necessary to construct a model for the transcriptional process is to infer the relationship between the rate of transcription and the length of time spent in any transcriptional state, which is shown in Figure 4.23. It can, thus, be seen that the duration of a transcriptional state appears to be only weakly dependent on the level of transcription. It is to be expected that if one had longer periods of observations, a clear dependence may be uncovered, particularly, given the tendency for short bursts of high transcriptional activity in the immature tissue (Figure 4.10). However, since no clear pattern emerges from Figure 4.23, for the purpose of the subsequent analysis, we consider the rate of transcription to be independent of the time spent in any transcriptional state.

Parametric Transcriptional Process

The independence between the transcriptional state and waiting time enables us to decouple the transcriptional process into two independent processes, one modelling

the transition times between states and the other modelling the transition between transcriptional levels. These two processes can be parameterised by equations (4.3) and (4.4) respectively.

Parameterising the posterior profiles in this way not only allows interpretable comparisons to be made between the pooled datasets of different developmental stage but also enables one to construct a parametric model for the transcriptional process. Explicitly, let $\mathbb{P}(\tau, \beta^*|t, \beta)\ \mathrm{d}\tau\ \mathrm{d}\beta^*$ be the probability that a cell switches in the interval $(t+\tau, t+\tau+\ \mathrm{d}\tau)$ to a new state in $(\beta^*, \beta^*+\ \mathrm{d}\beta^*)$ given a cell is currently in state β at time t. Furthermore, let $q(\tau|t)\ \mathrm{d}\tau$ be the probability the next switch occurs in the interval $(t+\tau, t+\tau+\ \mathrm{d}\tau)$ and is defined by equation (4.3) and let $\omega(\log\beta^*|\log\beta)\ \mathrm{d}\beta^*$ be the probability the next state will lie in $(\beta^*, \beta^*+\ \mathrm{d}\beta^*)$ and is defined by equation (4.4) then,

$$\mathbb{P}(\tau, \beta^*|t, \beta)\ \mathrm{d}\tau\ \mathrm{d}\beta^* = q(\tau|t)\ \mathrm{d}\tau\, \omega(\log\beta^*|\log\beta)\ \mathrm{d}\beta^*.$$

Consequently, the transcriptional process can be modelled as a random jump process and given a full transcriptional profile, $\beta(t)$, the likelihood is given by,

$$\begin{aligned} f(\beta|\theta) &= \prod_{i=1}^{K} \mathbb{P}(\Delta_j, \beta(t_{i+1})|\beta(t_i)), \\ &= \prod_{i=1}^{K} q(\Delta_j)\, \omega(\log\beta_{t_{j+1}}|\log\beta_{t_j}), \end{aligned} \tag{4.5}$$

where K is the number of switches and $t_1, ..., t_K$ are the timings of each transcriptional switch and $\Delta_j := t_{j+1} - t_j$.

Under the additional assumption that the parameters of the transcriptional model are constant across each dataset, the joint likelihood for all the transcriptional profiles of all cells within a single pooled dataset is given by,

$$f\left(\beta^{(1)}, ..., \beta^{(N)}|\theta\right) = \prod_{s=1}^{N} \left[\prod_{i=1}^{K^{(s)}} \mathbb{P}\left(t_{i+1}^{(s)} - t_i^{(s)}, \beta_{t_{i+1}}^{(s)}|\beta_{t_i}^{(s)}\right) \right], \tag{4.6}$$

where $K^{(s)}$ is the number of switches for cell s.

This parametric form will be especially useful when extending our analyses into a spatial dimension in Part III.

4.6 Discussion

This chapter has focussed on the implementation of the stochastic switch model (SSM) for transcription to real single cell imaging data. We find the reduced computational cost of the LNA to be advantageous when applying the methods to large quantities of relatively large (e.g. long time series) data.

We have shown how further analyses of the posterior transcriptional profiles may give insight into the underlying mechanisms of gene expression. For instance, we see that the estimated active transcriptional rates occupy a wide range of values, indicating a continuous range of gene activity as opposed to a tissue-wide binary activity. Little can be said about the specific regulatory network such as the number and logics of interacting transcription factors, but evidence of refractory periods between transcriptional switches suggests the behaviour is mechanistic in the sense that it deviates from a random memoryless process. Moreover, the evidence pointing towards a continuous distribution of active transcriptional rates suggests a complex regulatory system in order to achieve these different transcriptional states.

Thus, the SSM provides an approach that is both flexible and scientifically interpretable. The natural hierarchical structure enables the differentiation of intrinsic variability and transcriptional switches. This has been exemplified through the Prolactin gene application, since little is known about the regulation and our posterior inference shows a clear dynamic switching regime for many different transcriptional levels. This is in contrast to assuming *a priori* a specific regulatory network, which to ensure model identifiability, often requires conditioning on gene activity (Tkačik and Walczak, 2011) an assumption that does not correctly model the intrinsic noise (Thomas et al., 2012).

The Prolactin gene provides a good example for modelling gene expression through stochastic processes with random transcriptional pulses as it exemplifies features found in many different genes (Suter et al., 2011).

Moreover, the parametric analysis of the posterior profiles can be used as a hypothesis generator for different regulatory and signalling mechanisms. This will be investigated further in Part III.

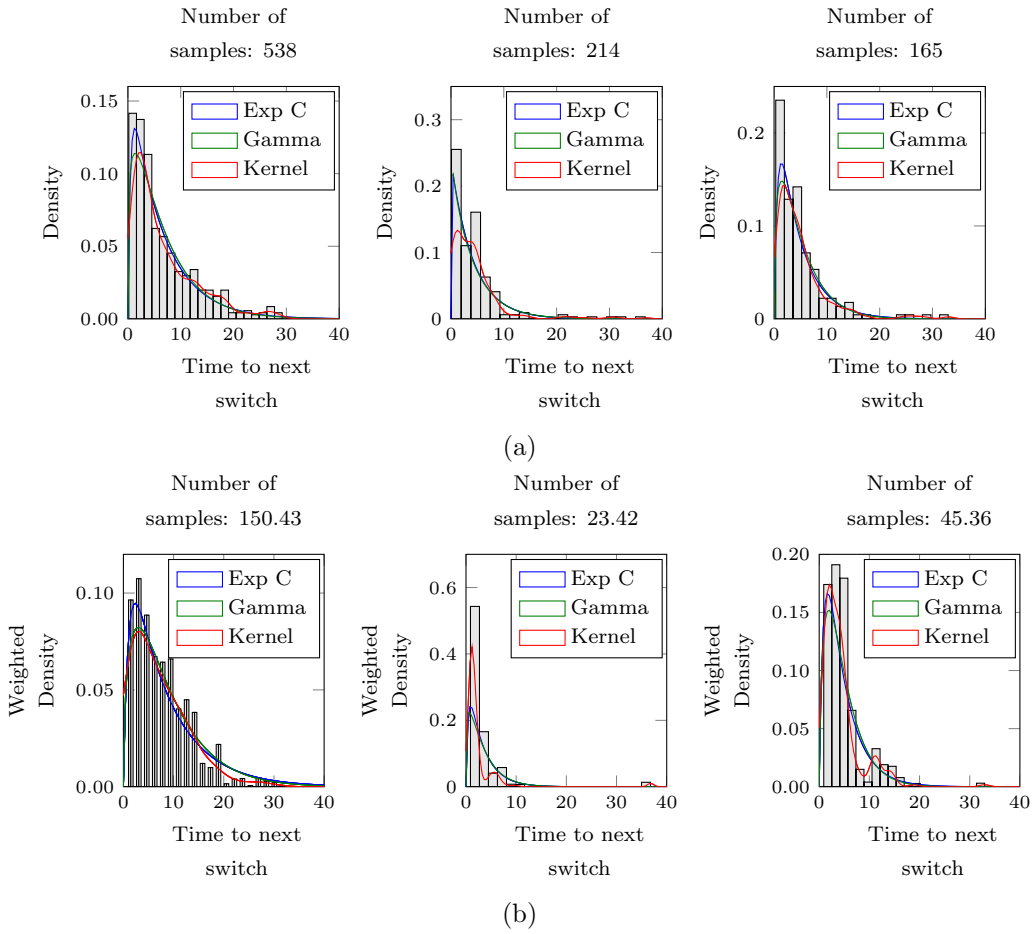


Figure 4.17: Posterior transcriptional waiting times between successive switches calculated a) based on the marginal transcriptional profiles for the pooled Adult (left), pooled P1.5 (middle) and pooled E18.5 (right) and b) based on the weighted conditional transcriptional profiles for the three pooled datasets. The blue line is given by fitting an Exponential convolution model, The green line is given by a Gamma fit and the red line is a kernel density fit to the data.

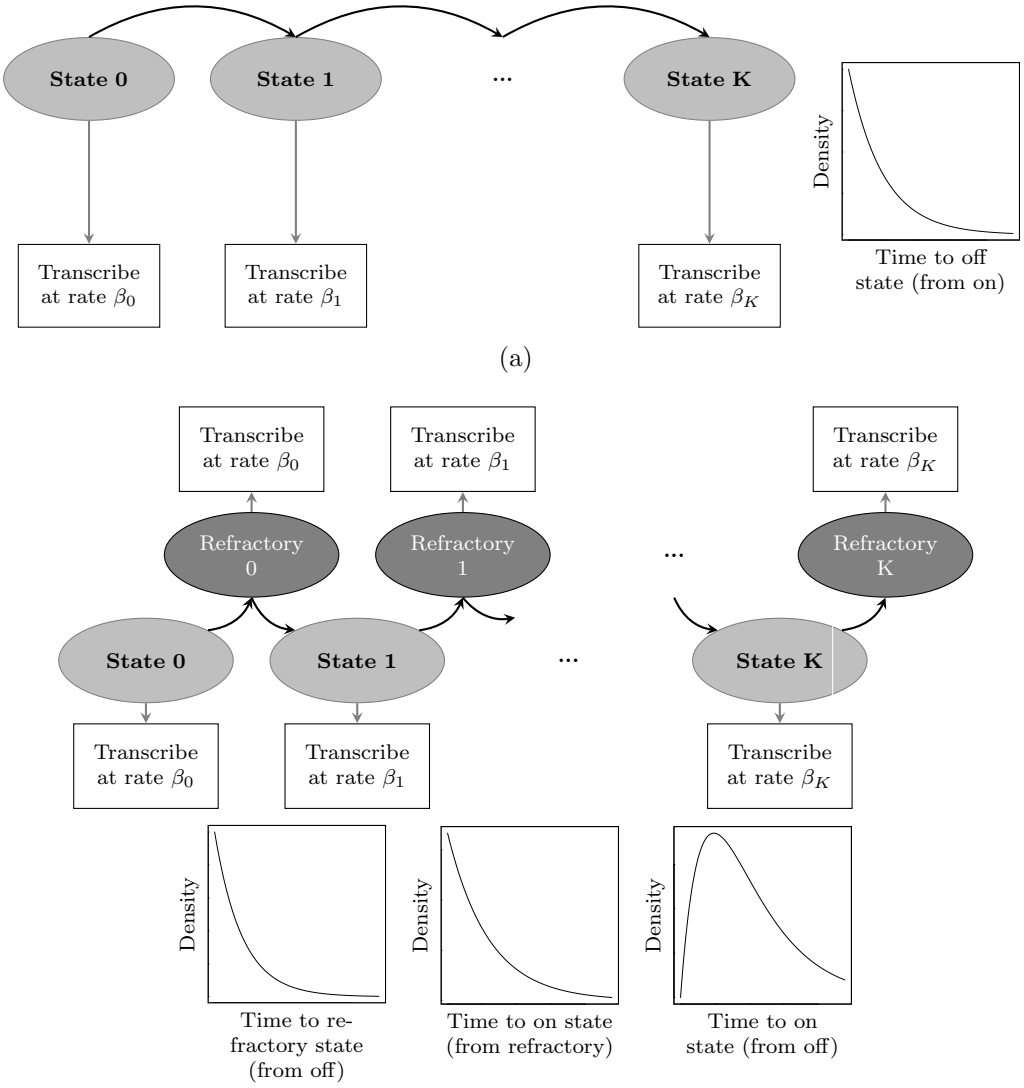


Figure 4.18: The multi-state switch model allows a gene to switch between any number of different states (a) with transcription occurring at any number of different rates. The transition times between these states follow an exponential distribution. Incorporating a refractory state (b) where transcription remains at the same rate as in the previous state, again with exponential waiting times between states, the overall transition density between different transcription states is given by the sum of exponentials.

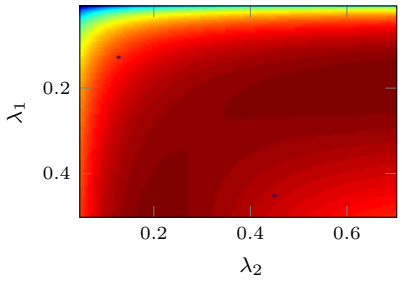


Figure 4.19: The likelihood surface of the two parameters of an exponential convolution model applied to the marginal waiting times of dataset A1.

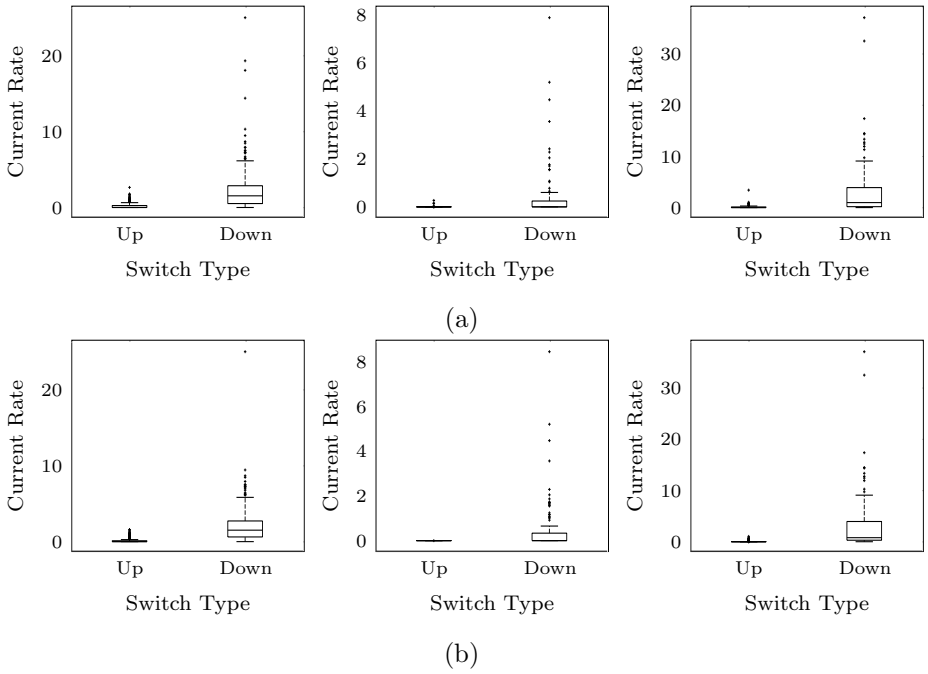


Figure 4.20: Boxplots showing how the type of next switch depends on the current transcriptional rate where a) is calculated under the marginal transcriptional process and b) under the weighted conditional transcriptional process. Left to right, the panels correspond to the pooled Adult, P1.5 and E18.5 datasets respectively.

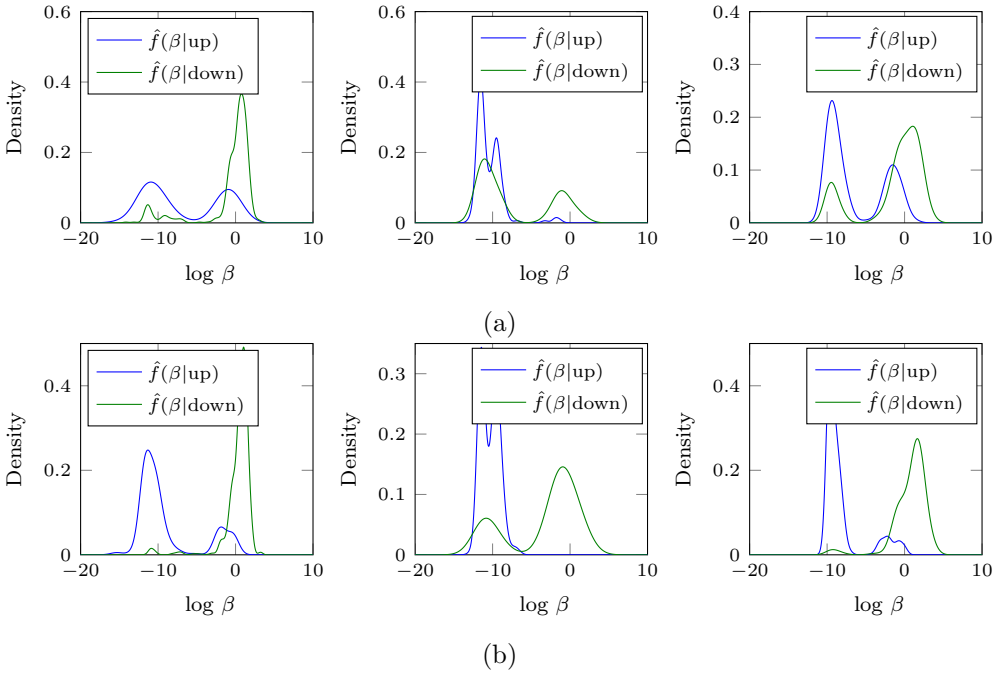


Figure 4.21: Density plots of how the current transcriptional rate changes with the type of next switch. Blue lines show the density of transcriptional rates conditioned on the next switch being an up switch and red lines indicate the density conditioned on a down switch where a) is calculated under the marginal transcriptional process and b) under the weighted conditional transcriptional process. Left to right, the panels correspond to the pooled Adult, P1.5 and E18.5 datasets respectively.

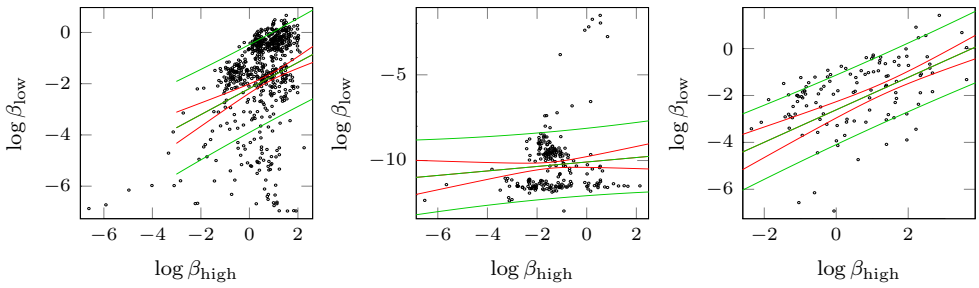


Figure 4.22: Linear regression fits between consecutive log transcriptional rates calculated under the weighted conditional transcriptional process. Left to right, the panels correspond to the pooled Adult, P1.5 and E18.5 datasets respectively. The red line shows the estimated mean response and associated 95% confidence intervals and the green line shows the corresponding 95% prediction intervals.

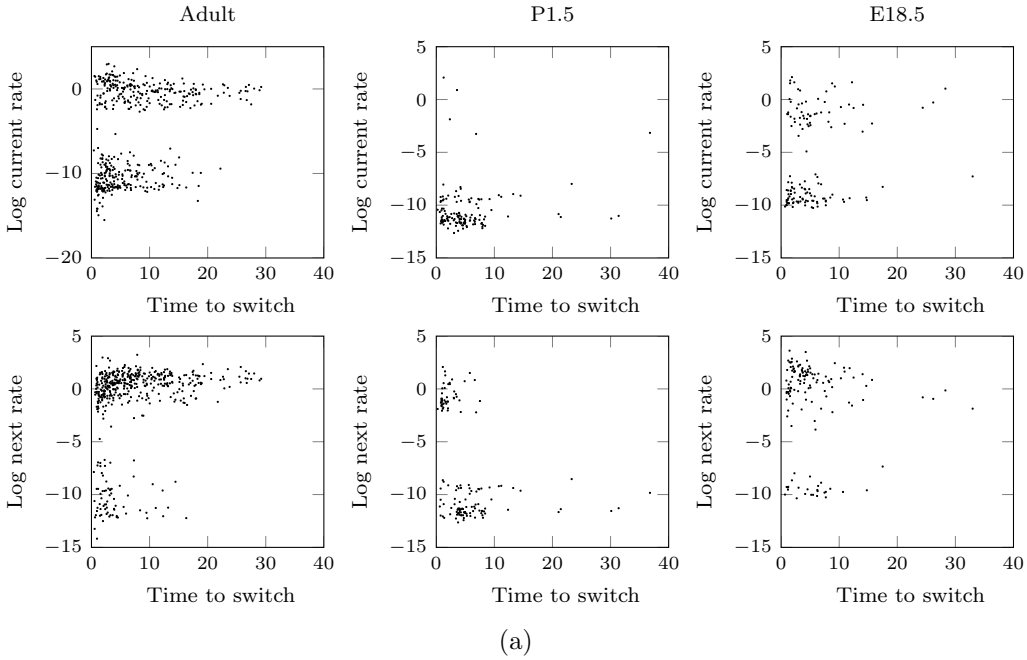


Figure 4.23: The relationship between the time spent in any transcriptional state and the level of transcription, both of the current state (top row) and the subsequent state (bottom row) where a) is calculated under the marginal transcriptional process and b) under the weighted conditional transcriptional process. Left to right, the panels correspond to the pooled Adult, P1.5 and E18.5 datasets respectively.

Part II

Spatial Organisation of Lactotroph Cells

CHAPTER 5

SPATIAL POINT PROCESSES

I have a hunch the most important reason we're going to space is not known now.

Burt Rutan

5.1 Introduction

As stated in the introduction, the data motivating this research provides a unique opportunity to study the spatial organisation of Prolactin producing cells (lactotrophs) within different experimental conditions. Primarily, these data are extracted from intact pituitary tissue in different stages of development and it is therefore desirable to characterise this spatial organisation both within a single dataset and between datasets over changing conditions.

There exist many types of mechanistic models that have been used for similar purposes and these are reviewed briefly in Chapter 1. However, due to the limited resolution of the data available and the fact that each dataset is an independent observation, a statistical analysis can be more appropriate. Specifically, we consider the application of spatial point processes to provide an acceptable framework for these analyses. Spatial point processes can be used to describe the spatial correlation between points, for example, to describe the level to which points show aggregation/clustering or repulsion/inhibition. The aim of this chapter is to provide the

necessary background into spatial point processes with the application to data given in Chapter 6. Consequently, the motivation for some of the models introduced in this chapter will become clear in the following chapter.

The remainder of this chapter is structured as follows. Section 5.2 introduces the notation and basic principles of spatial point processes, Section 5.3 describes the various techniques one can use for exploratory analysis including details of how these exploratory functions can be estimated. Section 5.4 introduces Cox processes whilst Section 5.5 goes on to discuss Gibbs processes. Finally, we present some examples of spatial point processes applied to biological data in Section 5.6.

5.2 Basic Principles

The majority of the background material for this section has been adapted from Møller and Waagepetersen (2004); Illian et al. (2008) and Baddeley and Turner (2005).

Informally, a spatial point process, $\mathbf{X}$, is a random countable subset of a space S. Throughout we will assume $S \subset \mathbb{R}^2$, and moreover let $W \subset S$ be the bounded observation window. Following Møller and Waagepetersen (2004), we restrict attention to point processes $\mathbf{X}$ whose realisations are locally finite subsets of S.

Definition 1. For any subset $\mathbf{x} \subset S$, let $n(\mathbf{x})$ denote the cardinality of $\mathbf{x}$, setting $n(\mathbf{x}) = \infty$ if $\mathbf{x}$ is not finite. Then $\mathbf{x}$ is said to be *locally finite* if $n(\mathbf{x}_B) < \infty$ whenever $B \subset S$ is bounded where $\mathbf{x}_B := \mathbf{x} \cap B$. Thus, $\mathbf{X}$ takes values in the space defined by,

$$N_{lf} = \{\mathbf{x} \subset S : n(\mathbf{x}_B) < \infty \, \forall \, B \subset S\}.$$

Let S be equipped with the Borel σ-algebra $\mathcal{B}$, and N_{lf} be equipped with the σ-algebra $\mathcal{N}_{lf} := \sigma\left(\{\mathbf{x} \in N_{lf} : n(\mathbf{x}_B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0\right)$, where $\mathcal{B}_0$ is the class of bounded Borel sets and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

Definition 2. Formally then, a *spatial point process*, $\mathbf{X}$ defined on S, is a measurable mapping defined on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and taking values in $(N_{lf}, \mathcal{N}_{lf})$. The *distribution* of $\mathbf{X}$ is given by,

$$F(D) := \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) \in D\}), \quad \text{for } D \in \mathcal{N}_{lf}.$$

If $\mathbf{x} = \{x_1, .., x_n\}$ is a realisation of the point process $\mathbf{X}$ then the number of points in

a point pattern $\mathbf{x}$ is given by $n(\mathbf{x}) = n$ and is a realisation from a random variable N. Consequently, knowing the distribution of N over all sets in $\mathcal{B}$ will uniquely define the point process. Moreover, it can be shown (Møller and Waagepetersen, 2004) that the probability distribution of $\mathbf{X}$ is uniquely determined by the *void probabilities* given by,

$$v(B) = \mathbb{P}\left(N(B) = 0\right), \quad B \in \mathcal{B}_0,$$

where $\{B \subset W : n(\mathbf{X} \cap B) = 0\}$ are the set of *void events*.

Definition 3. A *density*, f, for a point process $\mathbf{X}_1$ with respect to a process $\mathbf{X}_2$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are defined on the same space W, is given by,

$$\mathbb{P}\left(\mathbf{X}_1 \in B\right) = \mathbb{E}\left[\mathbb{I}\left[\mathbf{X}_2 \in B\right] f(\mathbf{X}_2)\right], \quad B \in \mathcal{B}. \tag{5.1}$$

The density, f, is a measurable function satisfying,

$$\mathbb{E}\left[f(\mathbf{X}_2)\right] = 1 \tag{5.2}$$

We shall often refer to the *unnormalised density*, h, given by,

$$f(\mathbf{x}) = c\, h(\mathbf{x}), \tag{5.3}$$

where the constant is given by, $c = \mathbb{E}\left[h(\mathbf{X}_2)\right]^{-1}$. As a consequence to (5.2), the unnormalised density must therefore always satisfy the following integrability condition,

$$\mathbb{E}\left[h(\mathbf{X}_2)\right] < \infty. \tag{5.4}$$

Definition 4. The *Papangelou conditional intensity* is defined to be,

$$\lambda(u, \mathbf{x}) = \frac{f(\mathbf{x} \cup \{u\})}{f(\mathbf{x} \setminus \{x\})}, \tag{5.5}$$

for $u \in S \setminus \mathbf{x}$ and can be interpreted as the conditional probability of finding a point at the location u given complete information about the process.

Definition 5. A point process is *stationary* or *homogeneous* if the process $\mathbf{X}$ and any translation of the process $\mathbf{X}$ are equal in distribution. Throughout, we shall use the terms stationary and homogeneous interchangeably. Moreover, a process is *isotropic* if any rotation of the process $\mathbf{X}$ is equal in distribution to the distribution

of $\mathbf{X}$.

Definition 6. The *intensity measure*, Λ, is given by,

$$\Lambda(B) := \int_B \lambda(u)\ \mathrm{d}u, \quad B \subset S,$$

for *intensity function*, λ, that is locally integrable ($\int_B \lambda(u)\ \mathrm{d}u < \infty$). Specifically, the intensity measure satisfies,

$$\mathbb{E}\left[N\left(\mathbf{X} \cap B\right)\right] = \Lambda\left(\mathbf{X} \cap B\right),$$

and if the process is stationary, the *intensity*, λ, is given by the mean number of points per unit area and satisfies,

$$\mathbb{E}\left[N\left(\mathbf{X} \cap B\right)\right] = \lambda\nu\left(\mathbf{X} \cap B\right),$$

where $\nu(B) = |B|$ is the volume of $B \subset S$.

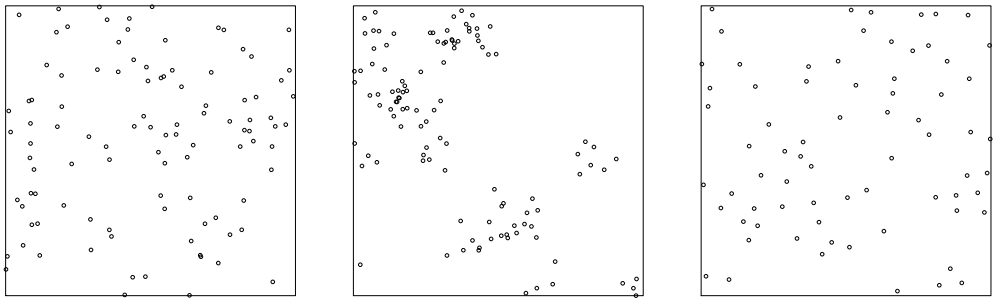
The "null" model of *complete spatial randomness* (CSR) has the properties that points are independent of each other and in addition, have the same propensity to be found at any location. Specifically, CSR is obtained under the homogeneous Poisson process with intensity λ, which satisfies the following two conditions,

1. The number of points $N(\mathbf{X} \cap B) \sim Pois(\Lambda(B))$ for any set $B \subset S$, where $\Lambda(B) = \lambda\,\nu(B)$, for a constant λ.
2. If B_1 and B_2 are disjoint sets, $N(\mathbf{X} \cap B_1)$ and $N(\mathbf{X} \cap B_2)$ are independent random variables.

A point process may deviate from complete spatial randomness in the following ways:

1. The underlying propensity of points is not constant over the window W, i.e. the process is *inhomogeneous*,
2. The positioning of points is not independent, i.e. the underlying distribution deviates from the Poisson assumption.

The unit ($\lambda = 1$) Poisson process, denoted by $\mathbf{Z}$, provides a reference model for an arbitrary point process $\mathbf{X}$. Specifically, using (5.1), a point process $\mathbf{X}$ is said to have



(a) (b) (c)

Figure 5.1: Realisations of spatial point processes where a) is a realisation from a Poisson process, b) a realisation of an inhomogeneous cluster/attractive process and c) a realisation of a non-Poisson regular/inhibitive process.

probability density function, f, with respect to the unit Poisson process if,

$$\mathbb{E}\left[g(\mathbf{X})\right] = \mathbb{E}\left[g(\mathbf{Z})f(\mathbf{Z})\right],$$

for some measurable functional g. Throughout, whenever we refer to the density of a point process, we specifically mean the density with respect to the unit Poisson process.

A single realisation of a Poisson process is shown in Figure 5.1 a). Figures 5.1 b) and c) show realisations of point processes deviating from CSR. Specifically, Figure 5.1 b) shows a realisation of an inhomogeneous cluster/attractive process and Figure 5.1 c) shows a realisation of a non-Poisson regular/inhibitive process.

5.3 Exploratory Analyses of Stationary Point Processes

There are a number of exploratory techniques one can use to explore certain features of a point process. These are typically related to the first and second moments or the inter-point distances of a stationary point process.

Ripley's *K-function* (Ripley, 1976) measures the average number of points found within a distance r from a typical point within a stationary point process and is defined by,

$$K(r) = \tfrac{1}{\lambda}\,\mathbb{E}\left[N(b(u,r) \setminus \{u\})|u \in \mathbf{X}\right], \quad \text{for } r \geq 0,$$

where $b(u,r)$ is a ball of radius r centred at u. Since a Poisson process describes

points that are randomly distributed in space, the associated K-function of a planar Poisson process is given by $K(r) = \pi r^2$. Moreover, if a process exhibits clustering, points are more likely to have close neighbours than in a Poisson process and $K(r) > \pi r^2$. Similarly, repulsive patterns will have fewer close neighbours than a Poisson process and $K(r) < \pi r^2$. A common transformation that stabilises the variance of the K-function is the *L-function*, given by,

$$L(r) = \sqrt{\frac{K(r)}{\pi}}, \quad \text{for } r \geq 0,$$

which under the Poisson process will be a straight line when plotted over r.

A closely related summary statistic is given by the *pair correlation function*, typically denoted by g, which is defined by,

$$g(r) = \frac{K'(r)}{2\pi r}, \quad \text{for } r \geq 0.$$

The pair correlation function can be interpreted as the probability of observing a pair of points separated by a distance r divided by the probability that a pair of randomly distributed points (i.e. from a Poisson process) are separated by the same distance r. Consequently, $g(r) > 1$ implies clustering and $g(r) < 1$ implies inhibition of a stationary point process.

Although the K-, L- and pair correlation functions are often considered as the most useful summary statistics (Illian et al., 2008), alternative measures can be defined through inter-point distances. For example, the *spherical contact distribution function*, which we shall denote by F, is defined by,

$$F(r) = \mathbb{P}\left[d(w, \mathbf{X}) \leq r\right], \quad \text{for } r \geq 0,$$

where $d(w, \mathbf{X}) = \min_{x \in \mathbf{X}} ||x - w||$, and w is an arbitrary reference location in W. This function is also referred to as the distribution function of the empty space distance. Similarly, the *nearest neighbour distribution function*, denoted by G, is given by,

$$G(r) = \mathbb{P}\left[d(u, \mathbf{X} \setminus \{u\}) \leq r | u \in \mathbf{X}\right], \quad \text{for } r \geq 0,$$

where $d(u, \mathbf{X} \setminus \{u\})$ is the distance from an arbitrary point $u \in \mathbf{X}$ to its nearest neighbour in $\mathbf{X}$.

A common transformation of the spherical contact and nearest neighbour distribu-

tion functions is the J-function (Lieshout and Baddeley, 1996) defined by,

$$J(r) = \frac{1 - G(r)}{1 - F(r)}, \quad \text{for } r \geq 0.$$

Consequently, the J-function compares the environment of a typical random point of the process to the environment of a fixed arbitrary point (Lieshout and Baddeley, 1996), which for a Poisson process will be equal, to give $J = 1$. $J < 1$, implies there is a higher density of points close to a typical point of the process when compared to an arbitrary point in space and consequently implies a clustering mechanism. Conversely, $J > 1$ implies there is a lower density of points close to a typical point of the process to result in an inhibitive or repulsive process.

Estimation

Estimation of these summary statistics is not always straightforward. Particular attention should be given to edge effects, as in practice stationary point processes are only observed on a finite window and the corresponding naïve estimators of the summary statistics will be biased. There are a number of edge correction methods available for the different summary statistics, including the, pointwise unbiased, reduced sample (or border) method and the, more efficient, spatial Kaplan Meier estimate (Baddeley and Gill, 1997). For example, to calculate the spherical contact distribution function the biased naïve estimator is given by,

$$\hat{F}(r) = \frac{1}{n} \sum_{i} \sum_{j \neq i} \mathbb{I}_{[||x_i - x_j|| \leq r]},$$

with unbiased reduced sample estimator given by,

$$\hat{F}(r) = \frac{1}{n_b} \sum_{i:b(x_i,r) \cup W} \sum_{j \neq i} \mathbb{I}_{[||x_i - x_j|| \leq r]},$$

where n_b is the number of points that lie further than a distance r away from the boundary. In contrast, the Kaplan-Meier estimate is calculated more efficiently, by throwing away fewer data and is obtained through product integration. Details can be found in Baddeley and Gill (1997) and as a simple illustration, the analogous one-dimensional survival analysis example is given by,

$$\hat{F}(t) = 1 - \prod_{s \leq t} \left(1 - \frac{\sum_i \sum_j ||x_i - x_j|| = s}{\sum_i \sum_j ||x_i - x_j|| \geq s}. \right),$$

Note that there are other edge correction methods available, specifically for the estimation of the K-function one can use a more efficient isotopic correction (Ohser, 1983; Ripley, 1991).

It is typical that one compares the estimated summary statistics of the observed point pattern to the summary statistics of known point patterns. For example, to compare the observed statistics to those that would be obtained from a completely random spatial process, one can generate Monte Carlo simulation envelopes of the summary statistics under a homogeneous Poisson process. Deviation from a completely random process, or equivalently deviation from a homogeneous Poisson process, can be detected via any discrepancy between the estimated summary statistics and the simulation envelopes.

An important caveat when using these summary statistics to explore an observed point process is that they are all derived under the assumption of stationarity. Consequently, a departure from the behaviour of a Poisson process may not imply spatial correlation but rather a departure from homogeneity. If a point process is known to be inhomogeneous, one may "homogenise" the process by removing the non-stationary trend and to check the behaviour of the residual process. If the residual process still shows evidence of departure from Poisson behaviour, it gives an indication of spatial dependence between points.

5.4 Cox Processes

The Poisson process provides a reference model for complete spatial randomness, but can be extended to incorporate spatial inhomogeneity. The resulting inhomogeneous Poisson process is then defined by the property that the number of points $N(\mathbf{X} \cap B) \sim Pois\left(\int_B \lambda(x)\ \mathrm{d}x\right)$ for any set $B \subset W$, where now the intensity function, λ, depends upon spatial location. As in the homogenous case, it is assumed that the number of points falling in disjoint sets are independent random variables. One can either estimate $\lambda(x)$ non-parametrically through a kernel density estimate or assume a parametric form. For example, if information of covariates were available (an example could be the capillary network of the tissue sample) one could include a regression term on these covariates (corresponding to the distance to the nearest capillary).

Generalising the inhomogeneous Poisson process further by allowing $\lambda(x) \sim \Lambda(x)$ to be random, yields the Cox process. I.e. conditional on an observation $\Lambda(x) = \lambda(x)$,

a single realisation of a Cox process is an inhomogeneous Poisson process. This enables a hierarchical framework for inference about Λ. Clearly, a single realisation of a Cox process is indistinguishable from an inhomogeneous Poisson process and it is often application specific as to which approach to use. Cox processes are very general and can model quite structured point processes depending on the amount of covariate information available. If covariates are not available, alternatives include the Log-Gaussian Cox process, which models $\exp(\Lambda(x))$ as a random Gaussian field or (Neyman-Scott) cluster processes where $\Lambda(x) = \sum_{c \in C} \alpha k(x - c)$, with C being a Poisson process and k a kernel.

The density of a Cox process restricted to a bounded region W, is given by,

$$f(\mathbf{x}) = \mathbb{E}\left[\exp\left(|W| - \int_W \Lambda(u)\ \mathrm{d}u\right) \prod_{u \in \mathbf{X}} \Lambda(u)\right],$$

which is rarely tractable.

Although Cox processes provide a very flexible framework for modelling point processes, due to the independence assumption of the underlying Poisson process, one is unable to explicitly model spatial dependence. This turns out to be important in our applications since the only information available about points is their location. Hypothetically, if more spatial information, such as capillary networks, were available, it may be possible to model cell locations as conditionally independent given these covariates. However, without this information, one requires an alternative approach to model spatial dependence.

5.5 Gibbs Processes

In contrast to Cox processes, which are defined through the intensity function, Gibbs (or Markov) processes are defined explicitly through the density function, which enables the incorporation of interactions between sets of points and hence allows one to model spatial dependence. This dependence structure can take many forms but we shall restrict ourselves to *pairwise interaction models*, where the unnormalised density is of the form,

$$h(\mathbf{x}) = \left[\prod_{i=1}^{n} b(x_i)\right] \left[\prod_{i<j} c(x_i, x_j)\right],$$

where $b : W \to [0, \infty)$ and $c : W \times W \to [0, \infty)$ are given functions. For a homogeneous process, the function b will be constant and the function c should be translation invariant in the sense that $c(x_i, x_j) = c(||x_i - x_j||)$ only depends on the distance between any two points. It is common to express c in terms of a pair potential function ϕ such that it is always translation invariant and satisfies,

$$\prod_{i<j} c(x_i, x_j) = \exp\left(-\sum_{i<j} \phi(||x_i - x_j||)\right). \tag{5.6}$$

The Papangelou conditional intensity for a pairwise interaction model is given by,

$$\lambda(u, \mathbf{x}) = b(u) \prod_i c(u, x_i). \tag{5.7}$$

The class of pairwise interaction models is very flexible and allows one to construct many different types of interaction. It is this ability to incorporate explicit dependence between points that give Gibbs models an advantage over Poisson type models. For instance, referring to equation (5.6), if $\phi(r) = 0$, then there is no interaction at a distance r. If $\phi(r) > 0$, it means that pairs of points a distance r apart have a very small contribution to the density and hence are less likely to occur. Conversely, for $\phi(r) < 0$, it means pairs of points that are a distance r apart, have a large contribution and therefore occur more frequently resulting in attractive behaviour at this distance.

Below, we present some relatively simple examples of Gibbs processes. Although this list is far from comprehensive, we shall see in the following section how they can be used to construct more complex processes.

Hardcore

A hardcore process has interaction defined by,

$$c(x_i, x_j) = \begin{cases} \infty, & ||x_i - x_j|| \leq r_0 \\ 1, & ||x_i - x_j|| > r_0, \end{cases} \tag{5.8}$$

where r_0 is the hardcore distance and no points are found closer than r_0 apart. Letting $r := ||x_i - x_j||$, the pair potential function, ϕ, of a hardcore process is given by, $\phi(r) = \infty$ for $r \leq r_0$ and $\phi(r) = 0$ for $r > r_0$. Consequently, conditional on all pairs of points exceeding a distance r_0 apart, the process exhibits neither attractive

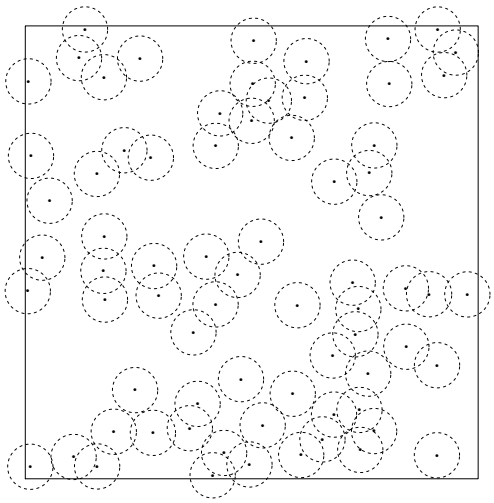


Figure 5.2: A realisation of a hardcore process, where each point has a minimum distance r_0 to any of its neighbours. This is shown by the dashed circle of radius r_0 about each point.

nor repulsive behaviour. This is a particularly useful property that allows one to model points as a random spatial point process whilst incorporating a non-zero physical size to the points, as shown in Figure 5.2.

The corresponding unnormalised density and conditional intensity are given by,

$$h(\mathbf{x}) = \left[\prod_{i=1}^{n} b(x_i)\right] \mathbb{I}\left[||x_i - x_j|| > r_0 \, \forall i \neq j\right], \tag{5.9}$$

$$\lambda(u, \mathbf{x}) = \begin{cases} b(u) & ||u - x_i|| > r_0, \\ 0, & \text{otherwise.} \end{cases} \tag{5.10}$$

Strauss

A Strauss process (Strauss, 1975) has interaction defined by,

$$c(x_i, x_j) = \begin{cases} \gamma, & ||x_i - x_j|| \leq r_1, \\ 1, & ||x_i - x_j|| > r_1, \end{cases} \tag{5.11}$$

where r_1 is the range of interaction and γ controls the intensity of points closer than r_1 apart. In order to satisfy the integrability condition (5.4), γ must be less than 1. Consequently, the corresponding pair potential function is given by $\phi(r) = -\log\gamma > 0$ for $r \leq r_1$, and $\phi(r) = 0$ for $r > r_1$. This means that the process

exhibits repulsive behaviour for points less than r_1 apart, since the contribution to the overall density will be small. Moreover, the Strauss process is only able to model repulsive behaviour since a clustering model with $\gamma > 1$ violates the integrability condition (5.4). The corresponding unnormalised density and conditional intensity are given by,

$$h(\mathbf{x}) = \left[\prod_{i=1}^{n} b(x_i)\right] \left[\gamma^{s(\mathbf{x})}\right], \tag{5.12}$$

$$\lambda(u, \mathbf{x}) = b(u)\gamma^{t(u,\mathbf{x})}, \tag{5.13}$$

respectively, where $s(\mathbf{x}) := \sum_{i<j} \mathbb{I}\left[||x_i - x_j|| < r_1\right]$ and $t(u, \mathbf{x}) := s(\mathbf{x} \cup \{u\}) - s(\mathbf{x})$.

Multiscale

A multiscale process (Penttinen, 1984) has interactions defined by,

$$c(x_i, x_j) = \begin{cases} \infty, & ||x_i - x_j|| \leq r_0, \\ \gamma_1, & r_0 < ||x_i - x_j|| \leq r_1, \\ \gamma_2, & r_1 < ||x_i - x_j|| \leq r_2, \\ \ldots & \\ \gamma_k, & r_{k-1} < ||x_i - x_j|| \leq r_k, \\ 1, & ||x_i - x_j|| > r_k, \end{cases} \tag{5.14}$$

where $\gamma_1, ..., \gamma_k < 1$ to satisfy (5.4) and the range of interaction is r_k. This process is able to model different types of spatial dependence at different spatial scales. For instance, consider the example satisfying,

$$c(x_i, x_j) = \begin{cases} \infty, & ||x_i - x_j|| \leq r_0, \\ \gamma_1, & r_0 < ||x_i - x_j|| \leq r_1, \\ \gamma_2, & r_1 < ||x_i - x_j|| \leq r_2, \\ 1, & ||x_i - x_j|| > r_2, \end{cases} \tag{5.15}$$

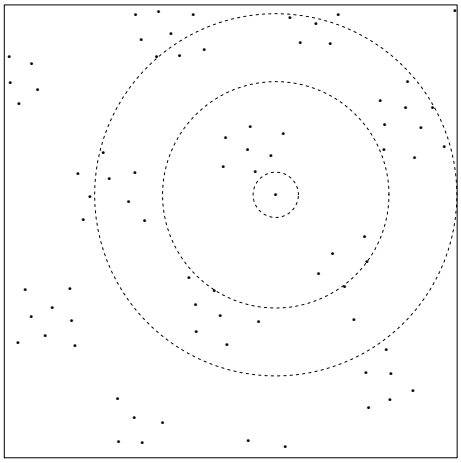


Figure 5.3: An illustration of how a two component multiscale process is constructed. Each point has a hardcore distance of r_0. Within the region $r_0 < r < r_1$ of any point, there will be a density of γ_1 points and in the region $r_1 < r < r_2$ there will be a density of γ_2 points, which in this simulation is greater than γ_1.

such that $\gamma_1, \gamma_2 < 1$ and $\gamma_1 < \gamma_2$. The corresponding pair potential function is then given by,

$$\phi(x_i, x_j) = \begin{cases} \infty, & ||x_i - x_j|| \leq r_0, \\ -\log \gamma_1, & r_0 < ||x_i - x_j|| \leq r_1, \\ -\log \gamma_2, & r_1 < ||x_i - x_j|| \leq r_2, \\ 0, & ||x_i - x_j|| > r_2, \end{cases} \tag{5.16}$$

and since $-\log \gamma_1 > -\log \gamma_2$, it means that although the process is repulsive, there will be a higher density of points within a distance r_1 to r_2 than in the range r_0 to r_1 as depicted in Figure 5.3.

It will be shown in the following section that this is an example of a much wider class of hybrid models that are able to incorporate different spatial dependencies at different spatial scales.

Geyer Saturation

A Geyer saturation process (Geyer, 1999) is defined by the density,

$$h(\mathbf{x}) = \left[\prod_{i=1}^{n} b(x_i)\right]\left[\prod_{i=1}^{n} \gamma^{\min\{q, t(x_i, \mathbf{x}\setminus x_i)\}}\right], \tag{5.17}$$

where $t(x_i, \mathbf{x} \setminus x_i) := \sum_{i \neq j} \mathbb{I}\left[||x_i - x_j|| < r_1\right]$ and $q \in \mathbb{N}$. The saturation parameter, q, ensures the density satisfies the integrability condition for all values $\gamma > 0$. The interpretation of the Geyer saturation model is similar to that of the Strauss process with an additional constraint to ensure mathematical tractability. The saturation parameter truncates the contribution to the overall density of each point. For example, if a point had 8 neighbours within a distance r, but the saturation threshold $q = 3$, the contribution of this point to the overall density would be the same as those points with only 3 neighbours. Consequently, the Geyer saturation process can capture both clustering and repulsive behaviour since the pair potential function, ϕ, can take any value in $(-\infty, \infty)$.

Area-Interaction

The area-interaction process (Widom and Rowlinson, 1970; Baddeley and Van Lieshout, 1995) is defined by the density,

$$h(\mathbf{x}) = \left[\prod_{i=1}^{n} b(x_i)\right]\left[\exp(-\kappa U(\mathbf{x}, r))\right], \tag{5.18}$$

where $U(\mathbf{x}, r) = |W \cap (\cup_i b(x_i, r))|$. This process captures the interaction through the so-called *area of influence*, consisting of the union of the radius r areas about each point. Consequently, κ can be viewed as a weight for the area of influence of each point. If $\kappa < 0$, the weight of the area of influence is large and hence the process is repulsive as the density becomes small. On the contrary for $\kappa > 0$, the weight of the area of influence is small resulting in a clustering process.

The area-interaction process may also be parameterised in the canonical scale-free form given by,

$$h(\mathbf{x}) = \left[\prod_{i=1}^{n} b(x_i)\right]\left[\eta^{-\frac{U(\mathbf{x}, r)}{2\pi r^2} + n}\right], \tag{5.19}$$

for $\eta > 0$. This reparameterisation allows an easier interpretation (Baddeley and

Turner, 2005), where the parameter η can be viewed as an interaction parameter and in contrast to the parameter κ of equation (5.18) is decoupled from the intensity function. For example under the parameterisation of equation (5.18), each point contributes a factor $b(x_i)(e^{\kappa})^{-\pi r^2}$ to the probability density. However, in the second parameterisation of equation (5.19), each point contributes a factor $b(x_i)$ to the density. In this reparameterisation, $\eta < 1$ corresponds to a clustered process and $\eta > 1$ to a clustered process. It should also be noted that the range of interaction between points is given by $2r$.

5.5.1 Hybrid Gibbs Processes

With the exception of the multiscale process, all models presented so far only incorporate spatial correlation at a single spatial scale. It is often of interest to model different correlation structures at different spatial scales since many factors may influence the spatial organisation. A general framework for achieving this has recently been proposed by Baddeley et al. (2013). In this paper, the authors introduce the notion of a hybrid point process defined through the unnormalised density, h. Specifically, let $h_1, h_2, ..., h_m$ be unnormalised densities. Then their *simple hybrid* is given by,

$$h(\mathbf{x}) = h_1(\mathbf{x})h_2(\mathbf{x}), ..., h_m(\mathbf{x}),$$

and the weighted hybrid is given by,

$$h(\mathbf{x}) = h_1(\mathbf{x})^{\kappa_1}h_2(\mathbf{x})^{\kappa_2}, ..., h_m(\mathbf{x})^{\kappa_m},$$

for $0 < \kappa_1, .., \kappa_m < \infty$.

In order the ensure the hybrid satisfies the integrability condition (5.4) and hence defines a point process, Baddeley et al. (2013) provide necessary conditions regarding the unnormalised density, h. In particular, if all of the individual components satisfy (5.4), the hybrid will be a Gibbs point process. However, the condition that *all* components are integrable is not strictly necessary. An interesting counter example is given in Baddeley et al. (2013) and presented below.

Hybrid Hardcore-Strauss

Let $h := h_1 h_2$, where h_1 is a Strauss density (5.12) with parameter γ, and h_2 be a hardcore density (5.9). The density $h(\mathbf{x})$ is given by,

$$h(\mathbf{x}) \propto \gamma^{s(\mathbf{x})} \mathbb{I}_{[||x_i - x_j|| > r_0, \forall i \neq j]},$$

and can be shown to be uniformly bounded and hence integrable for all $\gamma \in \mathbb{R}$.

This example is particularly interesting because although a Strauss process is only able to model inhibition (integrability condition ensures $\gamma < 1$), the hybrid of a Strauss and a hardcore process is able to model both inhibition and attraction since $\gamma \in \mathbb{R}$.

Hybrid models provide a very intuitive framework for modelling correlation at different spatial scales within a point process and moreover enable one to model both attraction and inhibition at these different scales. As such, we shall use these hybrid models extensively for analysing cell locations within different tissues.

5.5.2 Model Fitting

We consider here maximum likelihood approaches for estimating the parameters of a spatial point process, focussing our attention to the application of Gibbs processes. Specifically, the likelihood of a point process model is given by,

$$L(\theta) = f(\mathbf{x}|\theta) = h(\mathbf{x}|\theta) M(\theta)^{-1},$$

where h is the unnormalised density and $M(\theta) = \mathbb{E}[h(\mathbf{Z}|\theta)]$ is the normalising constant. Notice that we have explicitly denoted the dependence of the normalising constant on the unknown parameters, θ, of the point process. For example, under a Strauss process, θ will be the vector, $\theta := (r_0, r_1, \gamma, b)$. This normalising constant, $M(\theta)$, is typically intractable for Gibbs processes and moreover, the dependence on θ prevents the use of typical maximum likelihood methods. There are numerous variations of Monte Carlo approaches to maximum likelihood estimation and we present a brief overview of these methods with further details given in Møller and Waagepetersen (2004).

Monte Carlo Maximum Likelihood

Recall that a standard optimisation method for obtaining the MLE, is the Newton-Raphson algorithm, which is an iterative procedure producing a sequence of estimates of θ, $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, ...$, which will converge to the MLE, $\hat{\theta}$. The updating equation is given by,

$$\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} + u(\hat{\theta}^{(m)})j(\hat{\theta}^{(m)})^{-1}, \quad m = 0, 1...$$

where u is the score function and j is the observed Fisher information matrix as defined below. Since, the likelihood is only known up to a normalising constant, analytical expressions for u and j are unavailable. Consequently, one has to estimate them, which can be achieved through importance sampling. Let, $V_\theta(x) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log h(\mathbf{x}|\theta)$, and $\mathbb{E}_\theta[V_\theta(X)] = \frac{\mathrm{d}}{\mathrm{d}\theta} \log M(\theta)$. The score function, u, can then be expressed by,

$$\begin{aligned} u(\theta) &= \tfrac{\mathrm{d}}{\mathrm{d}\theta} \log f(\mathbf{x}|\theta) \\ &= V_\theta(x) - \tfrac{\mathrm{d}}{\mathrm{d}\theta} \log M(\theta), \\ &= V_\theta(x) - \mathbb{E}_\theta[V_\theta(X)]. \end{aligned}$$

Similarly, the observed Fisher information can be expressed in terms of V through,

$$\begin{aligned} j(\theta) &= -\tfrac{\mathrm{d}}{\mathrm{d}\theta} u(\theta) \\ &= -\tfrac{\mathrm{d}}{\mathrm{d}\theta} V_\theta(x) + \tfrac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log M(\theta), \\ &= -\tfrac{\mathrm{d}}{\mathrm{d}\theta} V_\theta(x) + \mathbb{E}_\theta \left[\tfrac{\mathrm{d}}{\mathrm{d}\theta} V_\theta(X) \right] + \mathrm{Var}_\theta[V_\theta(X)]. \end{aligned}$$

In order to obtain estimates of the score function and observed Fisher information matrix, consider the log-likelihood ratio for some fixed θ_0,

$$\log f(\mathbf{x}|\theta) - \log f(\mathbf{x}|\theta_0) = \log(h(\mathbf{x}|\theta)/h(\mathbf{x}|\theta_0)) - \log(M(\theta)/M(\theta_0)). \tag{5.20}$$

Due to the ratio of unknown normalising constants, equation (5.20) is not analytically available, but an importance sampling approximation is given by,

$$l_{\theta_0,n}(\theta) = \log(h(\mathbf{x}|\theta)/h(\mathbf{x}|\theta_0)) - \log \left[\tfrac{1}{n} \sum_{m=0}^{n-1} h_\theta(Y_m)/h_{\theta_0}(Y_m) \right], \tag{5.21}$$

where $Y_0, ..., Y_{n-1}$ is an MCMC sample with invariant density f_{θ_0}. Details of how to construct an appropriate MCMC sampler to simulate a spatial point process can

be found in Møller and Waagepetersen (2004) with particular attention given to the problem of estimating the ratio of unknown normalising constants.

Taking the first derivative of equation (5.21) then gives,

$$u_{\theta_0,n}(\theta) = V_\theta(x) - \mathbb{E}_{\theta,\theta_0,n}[V_\theta(X)],$$

the importance sampling approximation of the score function. Similarly, taking the second derivative of equation (5.21), one obtains the importance sampling approximation of the Fisher information matrix, $j_{\theta_0,n}(\theta)$. Consequently, a Newton-Raphson algorithm of the from,

$$\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} + u_{\theta_0,n}(\hat{\theta}^{(m)}) j_{\theta_0,n}(\hat{\theta}^{(m)})^{-1}, \quad m = 0, 1...$$

where $u_{\theta_0,n}$ and $j_{\theta_0,n}$ are the importance sampling approximations of $u(\theta)$ and $j(\theta)$ respectively can be used to obtain a maximum, $\hat{\theta}$, for an approximate log-likelihood.

It is noted in Møller and Waagepetersen (2004) that the above importance sampling procedure is only useful for $\hat{\theta}^{(m)}$ sufficiently close to θ_0. Consequently, when $\hat{\theta}^{(m)}$ deviates from θ_0 by a predefined distance, a new MCMC sample is required (where $\theta_0 \mapsto \hat{\theta}^{(m)}$) to generate new importance sampling approximations of $u(\theta)$ and $j(\theta)$.

Pseudo-likelihood Method

An alternative to Monte Carlo approximate maximum likelihood, is the pseudo-likelihood method of Besag (1975, 1977). This is defined through the Papangelou conditional intensities, under the assumption that the unnormalised density can be expressed in a log linear form,

$$h(\mathbf{x}) = \exp\left(\theta^{\mathrm{T}} S(\mathbf{x}) + B(\mathbf{x})\right), \tag{5.22}$$

for some functions of location, $S(\mathbf{x})$ and $B(\mathbf{x})$. Consequently, the Papangelou conditional intensity takes the form,

$$\lambda(u, \mathbf{x}) = \exp\left(\theta^{\mathrm{T}} T(u, \mathbf{x}) + C(u, \mathbf{x})\right), \tag{5.23}$$

where $T(u, \mathbf{x}) := S(\mathbf{x} \cup \{u\}) - S(\mathbf{x} \setminus \{u\})$ and $C(u, \mathbf{x}) := B(\mathbf{x} \cup \{u\}) - B(\mathbf{x} \setminus \{u\})$

Specifically, the pseudo-likelihood is given by,

$$PL(\theta) = \left[\prod_i \lambda(x_i, \mathbf{x}|\theta)\right] \exp\left(-\int_W \lambda(u, \mathbf{x}|\theta)\ \mathrm{d}u\right), \tag{5.24}$$

where λ is the Papangelou conditional intensity. This is derived (Jensen and Møller, 1991; Møller and Waagepetersen, 2004) in the limit as $i \to \infty$ of the quantity,

$$PL(\theta) = \exp(-|A|) \lim_{i\to\infty} \left[\prod_{j=1}^{m_i} f(\mathbf{x}_{A_{ij}}|\mathbf{x}_{W\setminus A_{ij}})\right], \tag{5.25}$$

where $\{A_{ij} : j = 1, ..., m_i\}$, $i = 1, ...$ are nested subdivisions of $A \subset W$ such that,

$$m_i \to \infty \quad \text{and} \quad m_i \left[\max_{1\le j\le m_i} |A_j|\right]^2 \to 0 \quad \text{as} \quad i \to \infty,$$

and f is the density of the point process.

Intuitively then, the pseudo-likelihood can be viewed as an infinite product of infinitesimal conditional probabilities (Baddeley and Turner, 2000) and up to a constant of $\exp(|W|)$, resembles the standard likelihood decomposition,

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\mathbf{x} \setminus \{x_i\}, \theta), \quad \text{for } \mathbf{X} \in W.$$

Example. *The Inhomogeneous Poisson Process.* The inhomogeneous Poisson process has a tractable likelihood and is given by,

$$f(\mathbf{x}|\theta) = \left[\prod_{i=1}^{n} \lambda(x_i)\right] \exp\left(|W| - \int_W \lambda(u)\ \mathrm{d}u\right), \tag{5.26}$$

where λ is both the Papangelou conditional intensity and the intensity function of the inhomogeneous process. The corresponding pseudo-likelihood is given by,

$$PL(\theta) = \left[\prod_{i=1}^{n} \lambda(x_i)\right] \exp\left(-\int_W \lambda(u)\ \mathrm{d}u\right), \tag{5.27}$$

and differs from (5.26) by a factor of $\exp(|W|)$.

Example. *A Pairwise Interaction Process.* Recall that the likelihood for a pairwise

interaction process is given by,

$$f(\mathbf{x}|\theta) = \left[\prod_{i=1}^{n} b(x_i)\right]\left[\prod_{i<j} c(x_i, x_j)\right] M(\theta)^{-1},$$

where $M(\theta)$ is the unknown normalising constant. Using equation (5.24), the corresponding pseudo-likelihood is given by,

$$PL(\theta) = \left[\prod_{i=1}^{n} b(x_i)\right]\left[\prod_{i<j} c(x_i, x_j)\right] \exp\left(-\int_W b(u) \prod_{j=1}^{n} c(u, x_j)\ \mathrm{d}u\right),$$

where the intractable normalising constant has been replaced by an exponential integral in the same vein as for a Poisson process (Baddeley and Turner, 2000).

Although the pseudo-likelihood is an approximation to the true likelihood, taking the first derivative, one obtains the following unbiased estimating function for θ, the pseudo-score,

$$PU(\mathbf{x}|\theta) = \sum_i T(x_i, \mathbf{x}) - \int_W T(u, \mathbf{x})\lambda(u, \mathbf{x}|\theta)\ \mathrm{d}u, \tag{5.28}$$

where $T(u, \mathbf{x}) := S(\mathbf{x} \cup \{u\}) - S(\mathbf{x} \setminus \{u\})$. Thus, solving $PU(\mathbf{x}|\theta) = 0$ will give an unbiased estimate for θ.

The pseudo-likelihood, relies on being able to express the unnormalised density in log-linear form, which may only be achievable when certain parameters (*nuisance parameters*) are fixed. For example, η is defined to be a nuisance parameter if the conditional intensities cannot be expressed in the log linear form of (5.23) and rather appears in the nonlinear form (Baddeley et al., 2013) below,

$$h(\mathbf{x}|\theta, \eta) = \exp\left(\theta^{\mathrm{T}} S(\mathbf{x}, \eta) + B(\mathbf{x}, \eta)\right). \tag{5.29}$$

In practice, one can obtain approximate estimates of the nuisance parameters by maximising the profile pseudo-likelihood given by,

$$PPL(\eta) = \max_{\theta} \log PL(\theta, \eta). \tag{5.30}$$

Then conditional on these estimates $\hat{\eta} := \arg\max_\eta PPL(\eta)$, the remaining parameters, θ, can be estimated by maximum pseudo-likelihood methods.

In the following chapter, we will restrict ourselves to basing inference through the

maximisation of the pseudo-likelihood and profile pseudo-likelihood for nuisance parameters as described in Baddeley et al. (2013). It should be noted that Illian et al. (2008) recommend this methodology be used as a preliminary analysis that can be further refined through a Monte Carlo maximum likelihood approach.

5.6 Literature Examples

There are many examples of the application of spatial point processes and it is beyond the scope of this thesis to provide a thorough or even representative review of these applications. However, in this section, we have picked two examples of how spatial point processes can be used to gain insight into different biological processes and hence motivates why we choose to analyse the Prolactin data in this broad framework presented in the following chapter.

Example. *Amacrine Cells.* Diggle (1986, 2005) presented the analysis of the spatial organisation of different cell types within the retina. Specifically, the authors were interested in distinguishing between two hypotheses describing the emergence of the different cell types during development. In particular, the authors were interested in distinguishing whether the two cell types emerge from a single undifferentiated population or if in fact they develop independently of each other. Through exploratory analysis of the bivariate K-function, it was found that the two cell types would be reasonably well modelled by two independent point processes favouring the second biological hypothesis. The specific point process model used to describe the spatial organisation was a pairwise interaction Gibbs process incorporating an inhibitory component between points to reproduce the effect created by the non-zero physical size of cells. Through pseudo-likelihood Monte Carlo tests of the fitted point process model, it was found that a weak dependence did exist between the two cell types that was not captured in the K-function exploratory analysis. The main conclusions of the paper were that although the exploratory analysis gives a good summary of the data and can indicate which family of point process models are appropriate for the analysis of the data, more information can be obtained through a modelling and inferential approach.

Example. *Microscopic and Macroscopic Biological Image Data.* Fleischer et al. (2006) present the analysis of two quite different biological applications with similar aims for inference. The first application considers the analysis of cell nuclei structure, specifically, the distribution of centromeres in two cell types, whereas the second application analyses the distribution of root systems for two different tree species.

In both applications, the authors were interested in making comparisons between two different subgroups of data and also in the comparison to completely random spatial behaviour. This was achieved through the use of the summary statistics, reviewed in Section 5.3. For example, the L- and pair correlation functions were particularly useful in detecting attractive behaviour in the root dataset.

As in the previous example, the use of summary statistics motivated a particular point process model to describe the features detected in the exploratory analysis, which here consisted of a Matérn-cluster process (an example of a Cox process), which could account for attractive behaviour and was fitted to the root dataset. Moreover, for the cell nuclei dataset, the pair correlation function was used to show that one cell type occurred with a greater clustering effect at certain distances compared to a different cell type, motivating further biological hypotheses.

Other examples of the application of point process modelling includes the location of forest fires (Turner, 2009), point patterns of Western Australia plants (Illian et al., 2009), the positioning of herd dispersion in Kenya (Stein and Georgiadis, 2006) and many other applications ranging from the ecological to the archaeological. For each application, different models are assumed and indeed, there are many datasets that have been analysed through a variety of different point processes. As demonstrated by the two examples given here, the exploratory analysis through the use of summary statistics provides a starting point for a modelling framework. The following chapter, follows this approach by providing an in-depth analysis of various summary statistics for lactotroph location within pituitary tissue, which is further analysed in a specific modelling and inferential framework. By nature of the data available, this modelling approach remains descriptive and cannot give huge insights into the underlying biology. However, it can generate hypotheses that may motivate future experiments. Moreover, we demonstrate the types of data one would wish to collect in order to model the cell organisation in a more mechanistic or biological framework than can currently be achieved.

CHAPTER 6

APPLICATION TO SINGLE CELL IMAGING DATA

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

John Tukey

6.1 Introduction

Over the last decade there has been considerable interest in identifying the biological structure and mechanisms of signalling within the mammalian pituitary. The majority of the pituitary is made up of five cell types, which are defined by the hormones they secrete. Specifically, somatotrophs secrete growth hormone (GH), lactotrophs secrete Prolactin (PRL), gonadotrophs secrete both luteinizing and folliculostellate hormone (LH and FSH), corticotrophs secrete adrenocorticotropic hormone (ACTH) and finally thyrotrophs secrete thyroid stimulating hormone (TSH) (Le Tissier et al., 2012). These cells appear through cell differentiation at different stages of pituitary development with corticotrophs appearing first. It has been shown that both corticotroph and somatotroph cells first appear as isolated cells that form clusters, which then form strands of connected cells (here connected refers to the presence of direct cell-cell contact). The gonadotroph network appears a few days later in development and again forms strands of connected cells (Mollard et al., 2012). Although

the lactotroph network has not been analysed in such detail there is evidence of interaction between the networks of different cell types. For instance there is a preferential location of gonadotrophs and lactotrophs within the pituitary (Le Tissier et al., 2012). To further complicate the network structure it has been shown that the vasculature of the tissue plays a role in these cell networks (Le Tissier et al., 2012). An important aspect in the identification of these connected cell networks has been the use of 3D imaging techniques (Bonnefont et al., 2005) that revealed the connectivity of cells in 3D that had previously appeared disconnected (i.e. not touching) in two dimensional slices.

Despite obvious limitations in the data available, for instance the lack of 3D imaging and no information on the tissue vasculature, we would like to characterise the spatial distribution of lactotroph cells. In contrast to the networks identified for other cell types, we are not looking for networks of physically connected cells but rather networks of cells that may provide a route of cell communication and signalling. This is motivated by Harper et al. (2010) who identified evidence of cell synchronicity (of luminescence signalling) between cells in close regions and moreover, found that when cells were enzymatically disaggregated, no synchronisation was evident.

As stated in the introduction, we have a number of datasets obtained from animals in different stages of development with both a spatial and temporal resolution. The aim of this section is to provide a statistical description of the spatial distribution of lactotroph cells and specifically, we are interested in how the distribution changes as the pituitary matures. Microscope images are shown in the first row of Figure 6.1 for three adult male pituitary slices (A1-A3), two P1.5 pituitary samples (P1-P2) and a single E18.5 pituitary sample (E1). Unfortunately, image files are not currently available for the remaining adult and E18.5 datasets (A4 and E2, respectively). The temporal data analysed in Chapter 4 were obtained by tracking a sample of the single cells identified in these images over time. Specifically, for datasets A1-A3 and P1-P2, only a small region of tissue (consisting of approximately 100 cells) was tracked of an approximate area of $150 \times 150 \mu m$ compared to the total image field of $400 \times 400 \mu m$. A larger area ($240 \times 220 \mu m$) was tracked in E1 to again consist of approximately 100 cells and corresponded to the majority of the image field being tracked. In this chapter, rather than analysing only the tracked cells, we instead choose to analyse the entire image field. This provides both a larger sample of data and also a better representation of the image boundary.

In order to extract cell location we first performed image registration techniques that allowed us to remove any drift in tissue movement and to align the sequence of

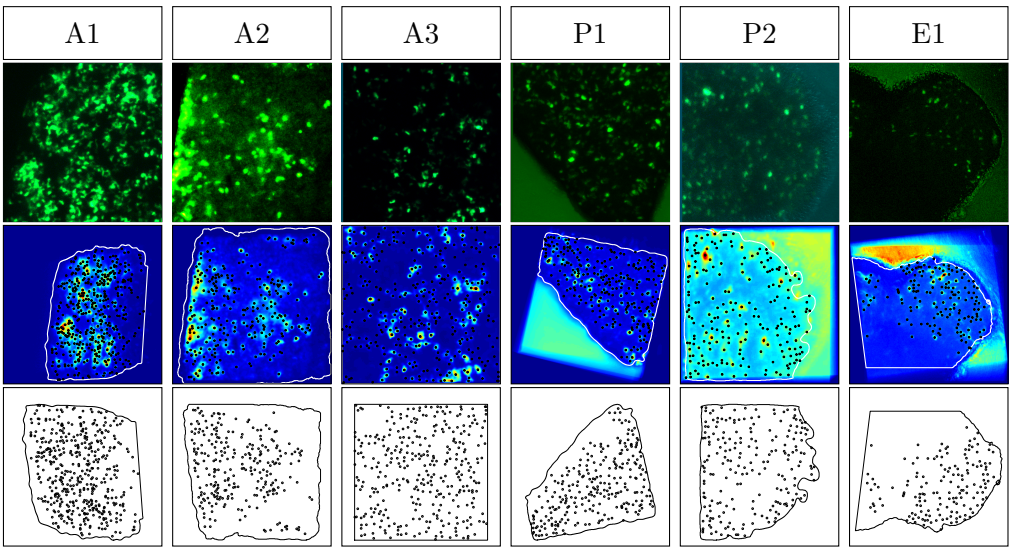


Figure 6.1: The first row shows an example time frame of the raw fluorescent imaging data for datasets A1-A3, P1-P2 and E1. The second row shows the combined registered image of the full time course where the colour intensity is given by the intensity of fluorescence levels. The black points indicate peaks of intensity and indicate cell location, shown in the third row. Note, the white boundaries shown in the second row are indicative of the tissue edge.

time frames. These frames were then combined into a single image and are shown in the second row of Figure 6.1. The cell locations were then extracted after a slight smoothing of the image through a simple peak detection algorithm and are shown by the black points in the second and third rows of Figure 6.1. It is these cell locations that can be viewed as a realisation of a spatial point process.

The remainder of this chapter is structured as follows, Section 6.2 will present a detailed exploratory analysis of the datasets, making use of the techniques outlined in Chapter 5. Section 6.3 will discuss how these observed point patterns can be parameterised by a particular spatial point process and Section 6.4 will discuss the interpretation of the fitted models.

6.2 Exploratory Analyses

In order to perform both exploratory and inferential analyses on these datasets, we have made extensive use of the **spatstat** package (Baddeley and Turner, 2005) in **R** (R Core Team, 2013).

The exploratory analysis presented in this section takes three forms. The first is to

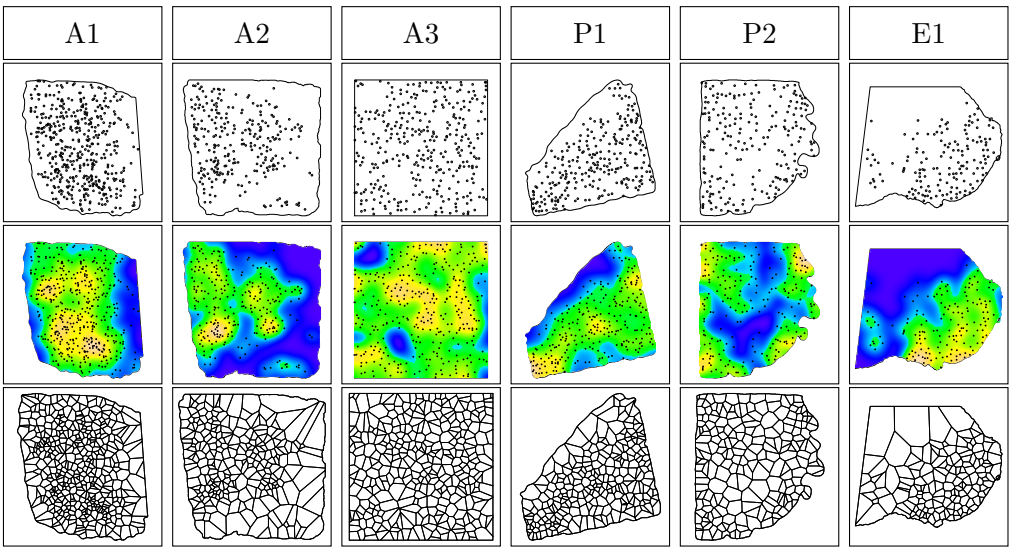


Figure 6.2: Spatial intensity analysis of the location of individual cells within datasets A1-A3, P1-P2 and E1. The first row shows the cell location with the second and third rows showing estimates of the intensity of points obtained from a kernel smoothing of point locations and the Voronoi tessellation, respectively.

analyse the intensity of the underlying point process in order to assess whether or not the process is stationary/homogeneous. More specifically, if there is a departure from stationarity it is of interest to capture the specific form of inhomogeneity. Secondly, we shall investigate the structure of different networks in the data as this may give useful characterisations of the connectivity of each tissue. Finally, we will analyse the summary statistics, defined in Chapter 5, which will help to identify particular parametric point processes that can model the observed data.

6.2.1 Intensity Analysis

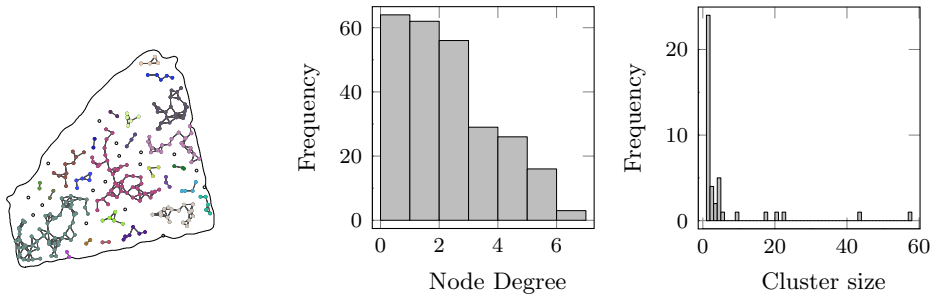
Assessing stationarity of a point process is difficult without knowing the underlying model. However, one can visually inspect the intensity of an observed point process to try to identify the presence or absence of a spatial trend. For our purposes, we have considered two different representations of the intensity function. The first is based on a kernel smoothing of the point process and the second on the Voronoi/Dirichlet tessellation. The kernel smoothing approach is highly dependent on the choice of bandwidth, here we have used the bandwidth that minimises the mean square error as defined in Berman and Diggle (1989). Since this bandwidth selection method has been shown to perform badly in some situations (Baddeley et al., 2013), we also look at the Voronoi tessellation. The Voronoi tessellation is

constructed by assigning each point a boundary, which is equidistant between any two neighbouring points. Consequently, for a homogeneous process the area contained within each boundary should be roughly equal over the whole tissue. For each dataset, both the kernel smoothing estimate of the intensity and the Voronoi tessellation are given in Figure 6.2. It is clear, particularly from the Voronoi tessellations that the only dataset that may be considered homogeneous is dataset A3. Specifically, A1-A2 and P1-P2 all show a clear inhibition very close to the boundaries that is likely to be an artefact of the data. Moreover, there is evidence in A1 and A2 and in dataset E1 of increased cell intensity towards the periphery of the tissue. This preferential location at the periphery of pituitary tissue has previously been identified experimentally (Featherstone et al., 2011; Harper et al., 2010). Dataset E1 gives the strongest indication of spatial inhomogeneity with a clear trend in the x-axis with preferential cell location at the tissue edge. In order to account for the spatial inhomogeneity evident in A1-A2, P1-P2 and E1, we will incorporate a polynomial spatial trend in the intensity function, which can be viewed as a proxy to the distance to the tissue boundary.

6.2.2 Network Analysis

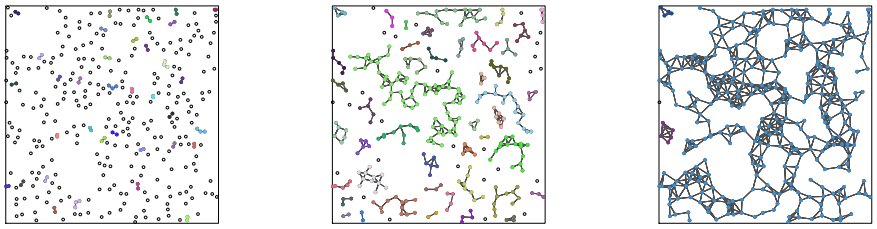
In the following chapter, we will investigate how (and if) the transcriptional dynamics of individual cells are correlated over space. In light of this, we are interested in identifying possible networks of communication through the spatial organisation of cells. To achieve this, we have looked at two different networks, the first is based on Euclidean distance and the second on geodesic distance. Despite the clear differences in spatial trend in the different datasets, the properties of these two networks remain relatively robust across the different tissues.

The Euclidean network is constructed by connecting any two cells with an edge if the distance between them is less than some predefined threshold d. For a fixed threshold d, we are interested in how the cells are connected. For example, Figure 6.3 a) shows the Euclidean network for dataset P1 for a threshold distance of $20\mu m$ with b) showing the distribution of node degree of the network and c) the size of each separate cluster in the network. Increasing the threshold d, we typically find the entire dataset becomes connected (i.e. one single cluster of cells) when $d \in (25\mu m, 48\mu m)$. Moreover, at the point the entire dataset becomes connected, the average node degree is around 4-5. Common to all datasets is the property that there is a very limited range of thresholds that produce a number of different clusters. For example, Figure 6.4 shows the Euclidean network for dataset A3 with threshold



(a) (b) (c)

Figure 6.3: Example of the Euclidean network for dataset P1 for a threshold distance of $20\mu m$ shown in a) with b) showing the distribution of node degree of the network and c) the size of each separate cluster in the network



(a) (b) (c)

Figure 6.4: Examples of the Euclidean network for dataset A3 with threshold a) $d = 10\mu m$ b) $d = 20\mu m$ and c) $d = 30\mu m$.

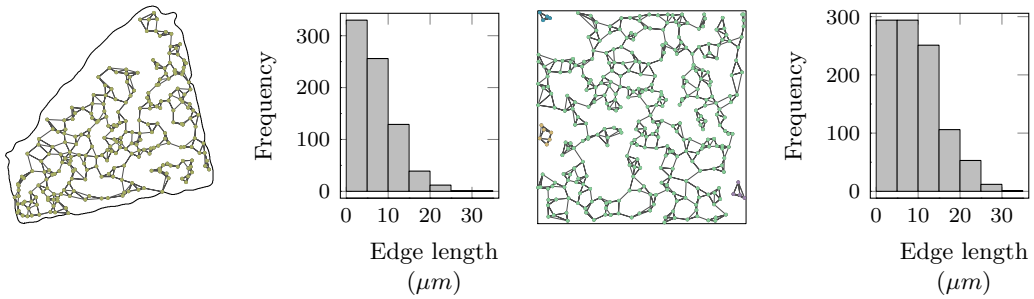
a) $d = 10\mu m$ - no connections b) $d = 20\mu m$ - clusters of cells appear and c) $d = 30\mu m$ - all cells are connected in one cluster. Table 6.1 gives the approximate range of threshold distances that give rise to this clustering behaviour. Thresholds below this range produce a dispersed network (50% of cells are completely disconnected) and thresholds above this range produce a fully connected network (90% of cells are contained within a single cluster). From this, all datasets seem to behave similarly although slightly more cell dispersion is apparent in the immature tissues.

In contrast, the geodesic network is constructed by connecting each cell to its k nearest neighbours. For a fixed k it is then of interest to examine the distribution of the edge length and the size of clusters that are produced. As for the Euclidean network, the different datasets behave similarly. In particular, we find that $k = 3$ connects the vast majority of the tissue in each of the different datasets and moreover, the distribution of edge lengths of these networks remains relatively consistent as shown in Figure 6.5 for datasets P1 and A3.

An interesting feature of the geodesic networks is the difference between the networks

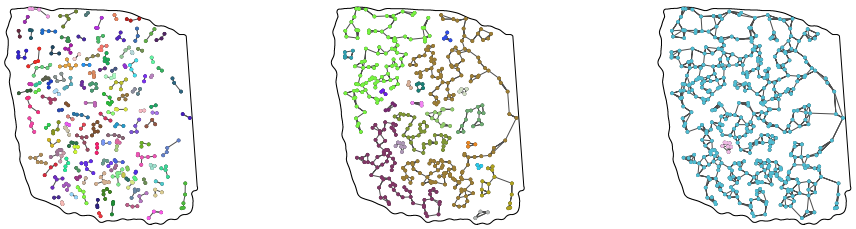
Dataset	Threshold Range (μm)
A1	(9.55, 23.05)
A2	(12.55, 35.05)
A3	(13.05, 29.05)
P1	(12.05, 25.55)
P2	(14.55, 31.05)
E1	(17.05, 47.55)

Table 6.1: Threshold ranges for Euclidean networks. Lower threshold is calculated when the number of disconnected cells is less than half the total number of cells. Upper threshold calculated when 90% of cells are connected by a single cluster. All ranges are given in μm.



(a) (b) (c) (d)

Figure 6.5: Geodesic network examples where each node is connected to its three nearest neighbours shown in a) for dataset P1 and c) for dataset A3. The corresponding distribution of edge lengths are shown in b) and d) respectively.



(a) (b) (c)

Figure 6.6: Geodesic networks for dataset A1 where each node is connected to its a) closest neighbour, b) two closest neighbours and c) three closest neighbours.

produced for $k = 1$ and $k = 2$. Specifically, for $k = 1$, the networks produced consist of a large number of small clusters whereas for $k = 2$, the networks generally consist of a small number of much larger clusters. This is exemplified in Figure 6.6 for dataset A1. These network properties are particularly relevant when it comes to modelling cell signalling. In particular, it is desirable to find at what distance cells signal and what the corresponding connectivity of a tissue is. It should be noted that the Euclidean networks will be more robust in the presence of missing data than the geodesic networks. Data may be missing for several reasons, for instance, although imaged for 48 hours, a lactotroph may not have been expressing Prolactin in this time and consequently won't have been imaged.

6.2.3 Summary Statistics

The final part of our exploratory analysis is to estimate those summary statistics given in Section 5.3. Some care needs to be taken when performing this analysis as we have previously identified possible deviations from spatial homogeneity. Figure 6.7 gives the observed L-, J- and pair correlation functions for each of the six datasets compared to a homogeneous Poisson process. Simulations were constructed from a homogeneous Poisson process and have been used to calculate both pointwise and simultaneous simulation envelopes. We show here in Figure 6.7, only the pointwise envelopes as it gives a better indication of the form of deviation from complete spatial randomness. Simultaneous simulation envelopes (not shown) enabled us to reject the hypothesis of complete spatial randomness as for each dataset, there was at least one summary statistic that did not lie within the significance bands.

For all datasets, there is a clear departure from complete spatial randomness, since the observed statistics deviate from the simulation envelopes. Looking specifically at the J-function, in most of the datasets, the observed statistic is larger than that obtained under a homogeneous Poisson process at short distances, indicative of repulsive behaviour. Analysing the L- and pair correlation functions, it can be seen that at larger distances, the observed statistic is larger than that obtained under a homogeneous Poisson process, which is indicative of attractive behaviour. However, in order to assess whether these departures from a homogeneous Poisson process were due to a departure in stationarity or a departure from independence, Figure 6.8 presents the summary statistics of each dataset compared to an inhomogeneous Poisson process with polynomial trend of order 3. If the departure from CSR were due only to the inhomogeneity of spatial trend (of polynomial order 3), the observed summary statistics should lie within the simultaneous simulation envelopes. Again,

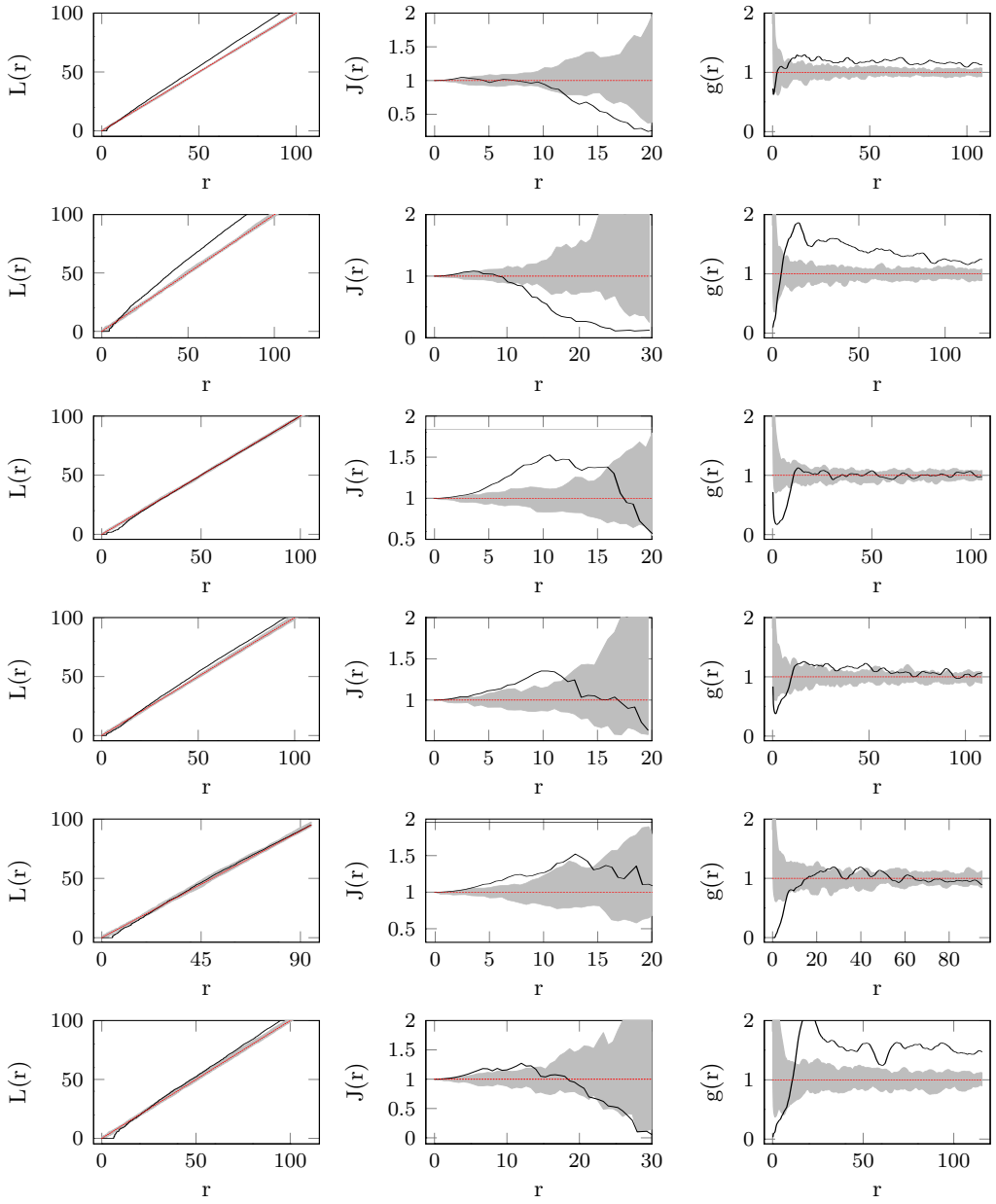


Figure 6.7: Observed L-, J- and pair correlation functions (black) for each dataset with 95% pointwise simulation envelopes obtained from a homogeneous Poisson process (grey region). The theoretical summary statistics of a homogeneous Poisson process are shown by the red dashed lines. Each row corresponds to the datasets A1-A3, P1-P2 and E1 respectively.

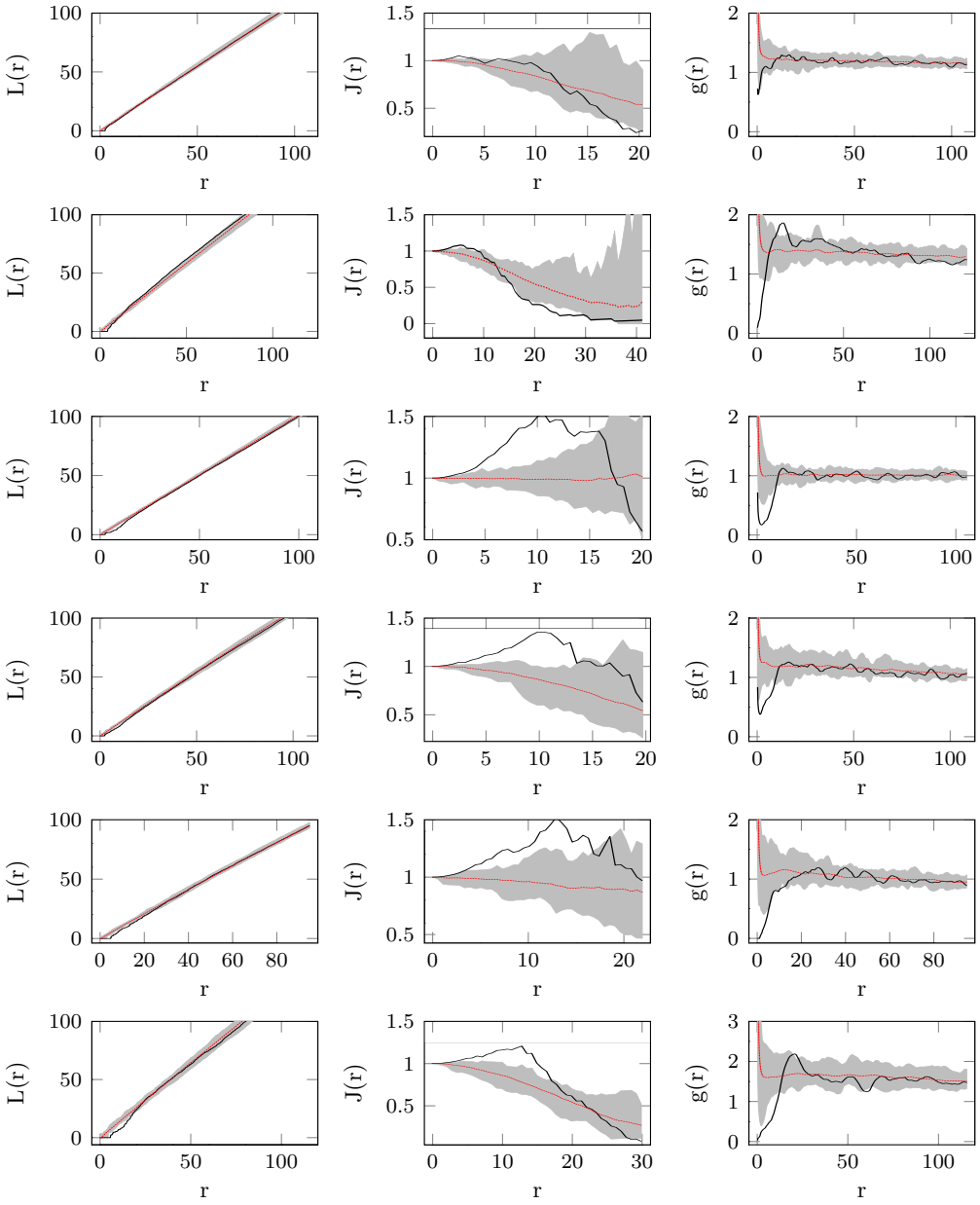


Figure 6.8: Observed L-, J- and pair correlation functions (black) for each dataset with 95% pointwise simulation envelopes obtained from an inhomogeneous Poisson process with intensity given by a polynomial trend of order 3 (grey region). The theoretical summary statistics of this inhomogeneous Poisson process are shown by the red dashed lines. Each row corresponds to the datasets A1-A3, P1-P2 and E1 respectively.

these are not shown, but there remained evidence of a departure of independence between point locations. Figure 6.8 shows the pointwise simulation envelopes for the summary statistics and it can be seen that the repulsive behaviour at short distances remains although the clustering at larger distances becomes less evident.

6.3 Point Process Modelling

6.3.1 Model Fitting

From our exploratory analyses, it appears that there are different spatial dependencies evident at different spatial scales in the separate tissues. We therefore choose to model these processes through a hybrid Gibbs process as outlined in Section 5.5 of the previous chapter. Model fitting proceeds in the same way as outlined in Baddeley et al. (2013), where components are added sequentially based on the behaviour of the residual process. We start by modelling the processes as a hardcore process to incorporate the property of a minimum point distance as a proxy for the non-zero physical size of a cell. Model parameters were estimated via the pseudo-likelihood procedure outlined in Section 5.5, using the **spatstat** package (Baddeley and Turner, 2005).

To illustrate the technique, we present the fitting procedure for dataset A2.

1. Fitting a Hardcore Strauss Model: First, we fit an inhomogeneous hardcore model with polynomial trend. The degree of the polynomial is chosen through a stepwise selection procedure based on the AIC. For dataset A2, this was optimised at a degree of 5. The coefficients of this polynomial were estimated via maximum pseudo-likelihood. In addition, the hardcore distance, r_0, is a nuisance parameter and is consequently fixed at the unbiased maximum pseudo-likelihood value of $\hat{r}_0 = \min_{i<j} ||x_i - x_j|| * n/(n+1) = 4.38\mu m$.

   The range of interaction of the Strauss component is also a nuisance parameter and is again fitted by maximising the profile pseudo-likelihood, shown in Figure 6.9 to give $\hat{r}_1 = 5.88\mu m$. The corresponding interaction parameter γ_1 is estimated to be 0.45 suggesting an inhibitive or regular process.

2. Analyse the Residual Process: The residual process can be analysed through the plots shown in Figure 6.10. Specifically, Figure 6.10 a) shows the behaviour of the theoretical summary statistics under a hardcore Strauss model with polynomial trend. It can be seen that the observed summary statistics all

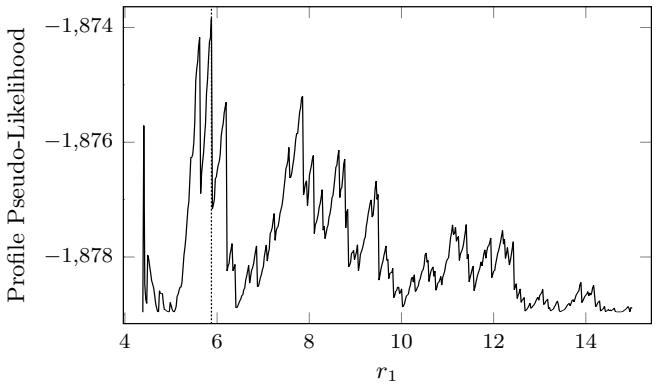


Figure 6.9: Profile pseudo-likelihood for the range of interaction in a hardcore Strauss model of dataset A2. Vertical line shows the maximum value.

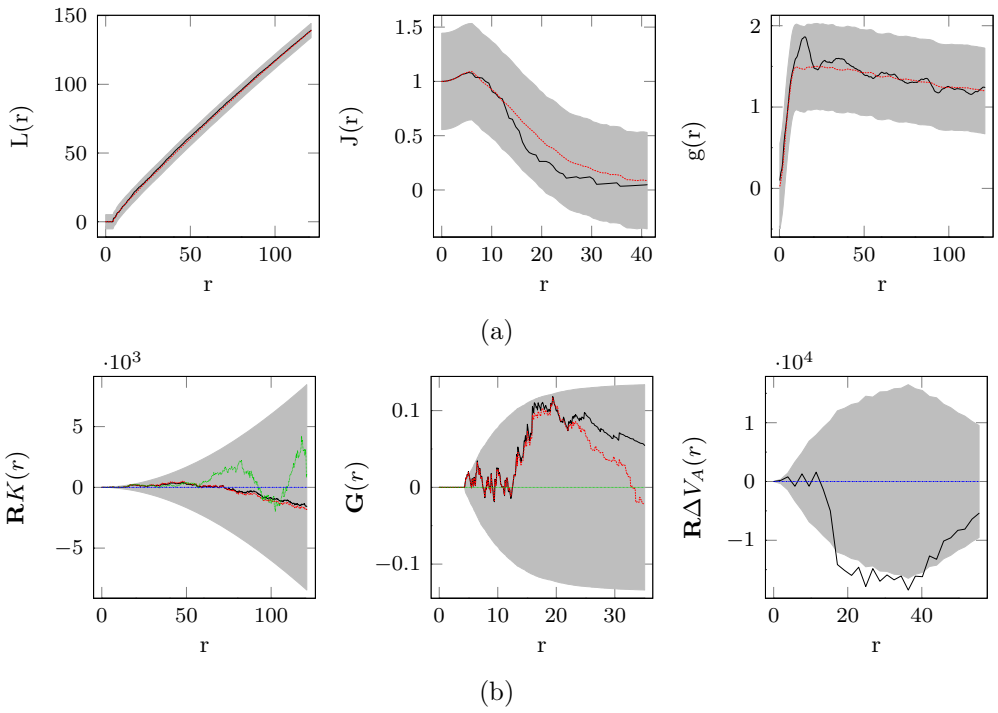


Figure 6.10: In a) are the summary statistics of dataset A2 shown by the black lines. The grey regions are given by the simultaneous 95% simulation envelopes calculated under the fitted hardcore Strauss model, centred about the mean shown in red. The three plots shown in b) correspond to a pseudo-score test of adding an additional Strauss, Geyer or Area-Interaction component to the model, respectively. The grey regions are 95% confidence envelopes, centred about the theoretical mean shown by the horizontal line. The black and red lines correspond to the empirical test function based on different border correction methods. Deviation from the grey region is indicative of favouring the alternative.

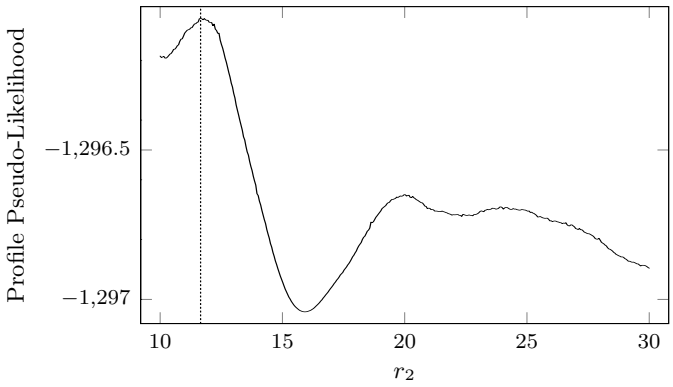


Figure 6.11: Profile pseudo-likelihood for the range of interaction in a hybrid hardcore Strauss Area-Interaction model of dataset A2. Vertical line shows the maximum value.

lie within the simulation envelopes calculated from the fitted model. This suggests the hardcore Strauss model to be a reasonable fit to the data. However, in Figure 6.10 b), there are three plots corresponding to three pseudo-score tests (Baddeley et al., 2011). The first corresponds to testing the addition of a Strauss component, the second to the addition of a Geyer component and the third to the addition of an Area-Interaction component. It can be seen that there is some evidence an additional Area-Interaction component should be included in the model.

3. Fitting the additional component: As with the Strauss component, the range of interaction, r_2, of the Area-Interaction component is estimated by maximising the profile pseudo-likelihood shown in Figure 6.11 and estimated to be $\hat{r}_2 = 2 \times 11.62\mu m$. Note, the factor of 2 comes from the parameterisation of the Area-Interaction process described in the previous chapter. The corresponding interaction parameter is estimated to be $\hat{\eta} = 1.18$ and revised estimate of the Strauss component is given by $\hat{\gamma} = 0.43$. this corresponds to a model with inhibitive behaviour at short distances and a clustering behaviour at interaction distances between $\hat{r}_1$ and $\hat{r}_2$.

4. Analyse Residual process: The residual process for the hybrid model is shown in Figure 6.12 and again there is no clear deviation between the summary statistics of the fitted model and the observed data. In addition, performing the same three tests as before, there is no suggestion that another component should be added.

We postpone the interpretation of the fitted model to Section 6.3.2 and first consider the assessment of model fit. The summary statistic analysis allows one to check that

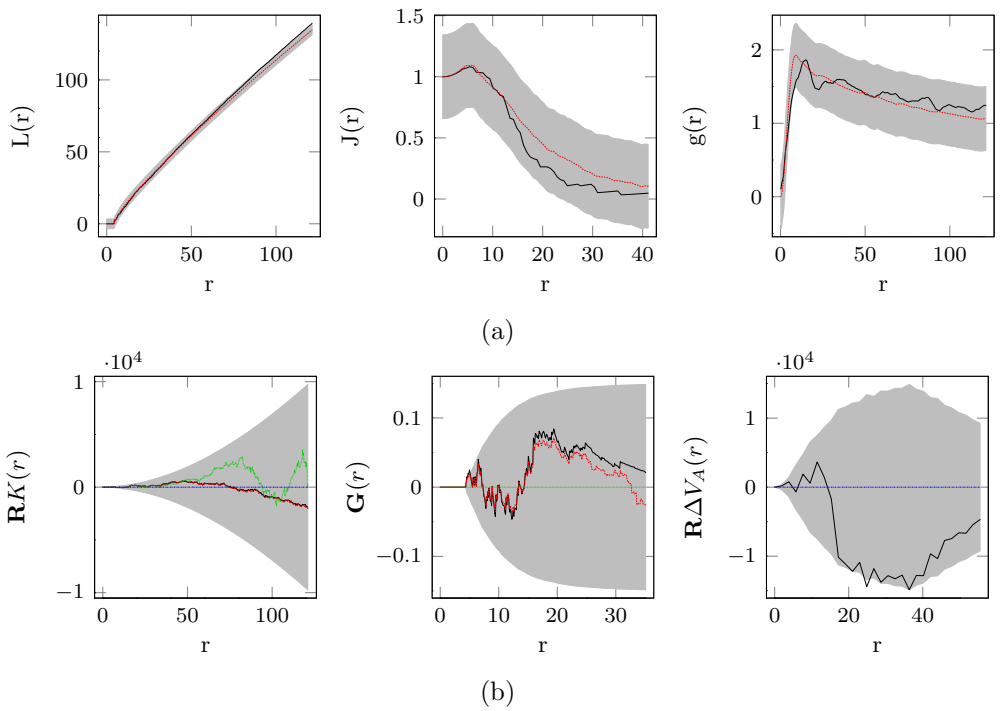
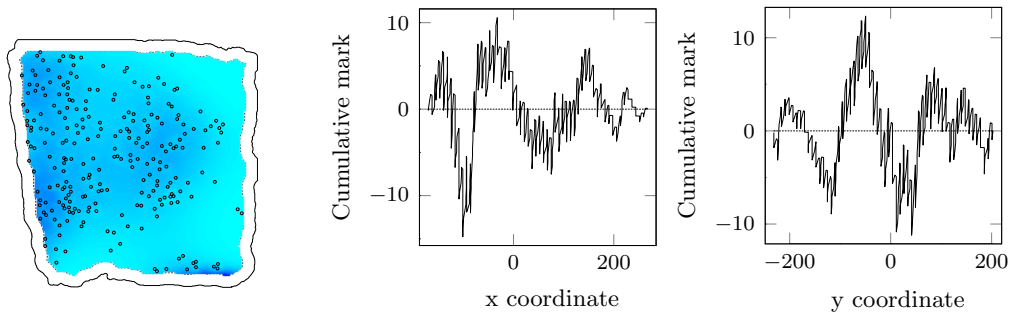


Figure 6.12: In a) are the summary statistics of dataset A2 shown by the black lines. The grey regions are given by the simultaneous 95% simulation envelopes calculated under the fitted hybrid hardcore Strauss and Area-Interaction model, centred about the mean shown in red. The three plots shown in b) correspond to a pseudo-score test of adding an additional Strauss, Geyer or Area-Interaction component to the model, respectively. The grey regions are 95% confidence envelopes, centred about the theoretical mean shown by the horizontal line. The black and red lines correspond to the empirical test function based on different border correction methods. Deviation from the grey region is indicative of favouring the alternative.

the fitted model reproduces certain features of the data but does not consider the fit as a whole. This is exemplified in the above example where both a hardcore Strauss model and a hybrid model reproduce the observed summary statistics. However, the AIC (albeit calculated with the pseudo-likelihood) shows a significant improvement with the hybrid model (a decrease from 3826 to 3653). One can additionally assess model fit through model residuals (Baddeley et al., 2005) defined by,

$$R(B) = n(\mathbf{x} \cap B) - \int_B \hat{\lambda}(u)\ \mathrm{d}u, \quad B \subset W, \tag{6.1}$$

where $\hat{\lambda}$ is the intensity of the fitted model. These residuals are useful for checking for misspecification of the spatial trend. An example for the fitted hybrid model to dataset A2 is shown in Figure 6.13, which contains three plots. For Gibbs models,



(a) (b) (c)

Figure 6.13: Residuals of the fitted hybrid point process to dataset A2, given by equation (6.1) are plotted in a) with b) and c) showing how the residuals vary with the x- and y- axis respectively.

these plots are not easy to interpret as significance bands cannot be produced. However, one can still look for "trend" in the x- and y- residuals, in Figure 6.13 b) and c), termed lurking variable plots, to identify misspecification of the intensity function. For example, Figure 6.13 may indicate some trend in the $x-$ and y-residuals and one might consider a different definition of the inhomogeneous spatial trend. Arguably, the more important misspecification to check, is the misspecification of point dependence. This is achieved through the analysis of the simulation envelopes of the summary statistics as outlined above.

6.3.2 Results

Applying the above procedure to each of the six datasets, we obtain the following set of hybrid Gibbs point process models. Parameter values for each of the interaction components for spatial dependence are presented for all datasets in Table 6.2.

Dataset A1

The final fitted model to dataset A1, is given by a hybrid of four components with a non-stationary polynomial trend of degree 5. Consequently, the unnormalised density of the final model is given by,

$$h(\mathbf{x}) = b(\mathbf{x})h_1(\mathbf{x})h_2(\mathbf{x})h_3(\mathbf{x})h_4(\mathbf{x}),$$

where the intensity function satisfies,

$$b(\mathbf{x}) = \prod_{i=1}^{n} b(x_i),$$

where $x_i := (x, y)$ and $b(x_i) = \text{poly}(x, y, 5)$ is a polynomial of degree 5. The four interaction components are given by,

1. Hardcore component:

$$h_1(\mathbf{x}) = \prod_{i<j} c_1(x_i, x_j), \qquad c_1(x_i, x_j) = \begin{cases} 0 & \text{if } ||x_i - x_j|| < r_0, \\ 1 & \text{if } ||x_i - x_j|| \geq r_0. \end{cases}$$

2. Strauss component:

$$h_2(\mathbf{x}) = \prod_{i<j} c_2(x_i, x_j), \qquad c_2(x_i, x_j) = \begin{cases} \gamma_2 & \text{if } ||x_i - x_j|| < r_1, \\ 1 & \text{if } ||x_i - x_j|| \geq r_1. \end{cases}$$

3. Geyer-Saturation component:

$$h_3(\mathbf{x}) = \prod_{i=1}^{n} \gamma_3^{\min\{q, t(\mathbf{x} \setminus x_i)\}},$$

   where $t(x_i, \mathbf{x} \setminus x_i) := \sum_{i \neq j} \mathbb{I}\left[||x_i - x_j|| < r_2\right]$.

4. Area-Interaction component:

$$h_4(\mathbf{x}) = \eta^{-\frac{U(x, r_3)}{2\pi r_3^2} + n},$$

   where $U(\mathbf{x}, r_3) = |W \cap (\cup_i b(x_i, r_3))|$.

We shall see that this hybrid model differs from the others in the Strauss component. Specifically, it is estimated as a cluster process (Table 6.2). However, since the range of interaction, $\hat{r}_1 = 3.83\mu m$ is much smaller than the typical cell size (estimated independently to be approximately $13.18\mu m$), this feature is likely to be a result of the measurement process. Specifically, this image file suffered from considerable image saturation, which means the microscope was unable to detect changes in fluorescence above a certain value. The result of this is a much smoother fluorescence image making it difficult to isolate individual cells within a certain distance, particularly when accounting for image drift. Having said this, beyond a certain radius,

we can still assume the measurements to be reliable and consequently components 3 and 4 are still representative of the dataset. In particular, the interaction function of component 4 (at a distance $\hat{r}_3 = 2 \times 18.27\mu m$) is estimated to be greater than 1 and hence a clustering process. Note that the estimated interaction parameters for all models are given in Table 6.2.

Dataset A2

The final hybrid model fitted to dataset A2 is given by three components with a non-stationary polynomial trend of degree 5. Consequently, the unnormalised density of the final model is given by,

$$h(\mathbf{x}) = b(\mathbf{x})h_1(\mathbf{x})h_2(\mathbf{x})h_3(\mathbf{x}),$$

where the intensity function satisfies,

$$b(\mathbf{x}) = \prod_{i=1}^{n} b(x_i),$$

where $x_i := (x, y)$ and $b(x_i) = \text{poly}(x, y, 5)$ is a polynomial of degree 5. The three interaction components are given by,

1. Hardcore component:

$$h_1(\mathbf{x}) = \prod_{i<j} c_1(x_i, x_j), \qquad c_1(x_i, x_j) = \begin{cases} 0 & \text{if } ||x_i - x_j|| < r_0, \\ 1 & \text{if } ||x_i - x_j|| \geq r_0. \end{cases}$$

2. Strauss component:

$$h_2(\mathbf{x}) = \prod_{i<j} c_2(x_i, x_j), \qquad c_2(x_i, x_j) = \begin{cases} \gamma_2 & \text{if } ||x_i - x_j|| < r_1, \\ 1 & \text{if } ||x_i - x_j|| \geq r_1. \end{cases}$$

3. Area-Interaction component:

$$h_3(\mathbf{x}) = \eta^{-\frac{U(x,r_2)}{2\pi r_2^2}+n},$$

where $U(\mathbf{x}, r_2) = |W \cap (\cup_i b(x_i, r_2))|$.

The second component, the Strauss process, is estimated to have an inhibitory effect implying few points are found within a distance $\hat{r}_1 = 5.88\mu m$ of each other. Since $5.88\mu m$ is smaller than the average cell size, this inhibitory effect may be the result of how the data are collected through a 2D projection. This feature is to some extent common to all datasets.

Similarly to dataset A1, the final component is estimated to have a clustering effect up to an interaction distance of $2 \times 11.65\mu m$.

Dataset A3

The final hybrid model fitted to dataset A3 is given by three components with a stationary trend. As conjectured through the intensity analysis of Section 6.2.1, we found that the optimal trend for this dataset to be constant. We hypothesise that this is due to the lack of a boundary effect as the tissue edge is not in the image field. Consequently, the unnormalised density of the final model is given by,

$$h(\mathbf{x}) = b(\mathbf{x})h_1(\mathbf{x})h_2(\mathbf{x})h_3(\mathbf{x}),$$

where the intensity function satisfies,

$$b(\mathbf{x}) = \prod_{i=1}^{n} b,$$

where b is a constant. The three interaction components are given by,

1. Hardcore component:

$$h_1(\mathbf{x}) = \prod_{i<j} c_1(x_i, x_j), \qquad c_1(x_i, x_j) = \begin{cases} 0 & \text{if } ||x_i - x_j|| < r_0, \\ 1 & \text{if } ||x_i - x_j|| \geq r_0. \end{cases}$$

2. Strauss component:

$$h_2(\mathbf{x}) = \prod_{i<j} c_2(x_i, x_j), \qquad c_2(x_i, x_j) = \begin{cases} \gamma_2 & \text{if } ||x_i - x_j|| < r_1, \\ 1 & \text{if } ||x_i - x_j|| \geq r_1. \end{cases}$$

3. Geyer-Saturation component:

$$h_3(\mathbf{x}) = \prod_{i=1}^{n} \gamma_3^{\min\{q, t(\mathbf{x}\setminus x_i)\}},$$

where $t(x_i, \mathbf{x} \setminus x_i) := \sum_{i \neq j} \mathbb{I}\left[||x_i - x_j|| < r_2\right]$.

Although parameterised through a Geyer-saturation process, dataset A3 behaved similarly to A2 with clustering seen in the third component up to an interaction range of $32.29\mu m$.

Dataset P1

The final hybrid model fitted to dataset P1 is given by three components with a non-stationary polynomial trend of degree 3. Consequently, the unnormalised density of the final model is given by,

$$h(\mathbf{x}) = b(\mathbf{x})h_1(\mathbf{x})h_2(\mathbf{x})h_3(\mathbf{x}),$$

where the intensity function satisfies,

$$b(\mathbf{x}) = \prod_{i=1}^{n} b(x_i),$$

where $x_i := (x, y)$ and $b(x_i) = \text{poly}(x, y, 3)$ is a polynomial of degree 3. The three interaction components are given by,

1. Hardcore component:

$$h_1(\mathbf{x}) = \prod_{i<j} c_1(x_i, x_j), \qquad c_1(x_i, x_j) = \begin{cases} 0 & \text{if } ||x_i - x_j|| < r_0, \\ 1 & \text{if } ||x_i - x_j|| \geq r_0. \end{cases}$$

2. Strauss component:

$$h_2(\mathbf{x}) = \prod_{i<j} c_2(x_i, x_j), \qquad c_2(x_i, x_j) = \begin{cases} \gamma_2 & \text{if } ||x_i - x_j|| < r_1, \\ 1 & \text{if } ||x_i - x_j|| \geq r_1. \end{cases}$$

3. Area-Interaction component:

$$h_3(\mathbf{x}) = \eta^{-\frac{U(x, r_2)}{2\pi r_2^2} + n},$$

where $U(\mathbf{x}, r_2) = |W \cap (\cup_i b(x_i, r_2))|$.

In contrast to the three adult datasets, the final component was estimated to have a regulatory effect, ensuring the process is inhibitive at all spatial scales.

Dataset P2

The final hybrid model fitted to dataset P2 is given by three components with a non-stationary polynomial trend of degree 2. Consequently, the unnormalised density of the final model is given by,

$$h(\mathbf{x}) = b(\mathbf{x})h_1(\mathbf{x})h_2(\mathbf{x})h_3(\mathbf{x}),$$

where the intensity function satisfies,

$$b(\mathbf{x}) = \prod_{i=1}^{n} b(x_i),$$

where $x_i := (x, y)$ and $b(x_i) = \text{poly}(x, y, 2)$ is a polynomial of degree 2. The three interaction components are given by,

1. Hardcore component:

$$h_1(\mathbf{x}) = \prod_{i<j} c_1(x_i, x_j), \qquad c_1(x_i, x_j) = \begin{cases} 0 & \text{if } ||x_i - x_j|| < r_0, \\ 1 & \text{if } ||x_i - x_j|| \geq r_0. \end{cases}$$

2. Strauss component:

$$h_2(\mathbf{x}) = \prod_{i<j} c_2(x_i, x_j), \qquad c_2(x_i, x_j) = \begin{cases} \gamma_2 & \text{if } ||x_i - x_j|| < r_1, \\ 1 & \text{if } ||x_i - x_j|| \geq r_1. \end{cases}$$

3. Area-Interaction component:

$$h_3(\mathbf{x}) = \eta^{-\frac{U(x,r_2)}{2\pi r_2^2}+n},$$

where $U(\mathbf{x}, r_2) = |W \cap (\cup_i b(x_i, r_2))|$.

The final component is again an Area-Interaction process, but in contrast to P1, was estimated to have a clustering effect albeit at a relatively small spatial scale of $2 \times 9.09\mu m$.

Dataset E1

The final hybrid model fitted to dataset E1 is given by two components with a non-stationary polynomial trend of degree 3. Consequently, the unnormalised density of the final model is given by,

$$h(\mathbf{x}) = b(\mathbf{x})h_1(\mathbf{x})h_2(\mathbf{x}),$$

where the intensity function satisfies,

$$b(\mathbf{x}) = \prod_{i=1}^{n} b(x_i),$$

where $x_i := (x, y)$ and $b(x_i) = \text{poly}(x, y, 3)$ is a polynomial of degree 3. The two interaction components are given by,

1. Hardcore component:

$$h_1(\mathbf{x}) = \prod_{i<j} c_1(x_i, x_j), \qquad c_1(x_i, x_j) = \begin{cases} 0 & \text{if } ||x_i - x_j|| < r_0, \\ 1 & \text{if } ||x_i - x_j|| \geq r_0. \end{cases}$$

2. Strauss component:

$$h_2(\mathbf{x}) = \prod_{i<j} c_2(x_i, x_j), \qquad c_2(x_i, x_j) = \begin{cases} \gamma_2 & \text{if } ||x_i - x_j|| < r_1, \\ 1 & \text{if } ||x_i - x_j|| \geq r_1. \end{cases}$$

As with all other datasets (excluding A1), the hardcore Strauss model was estimated to be inhibitive. No further components are estimated at larger ranges implying the process behaves like a random spatial point process conditional on points being further than $13.28\mu m$ apart.

Thus, bringing together the above analysis, we can start to draw some hypotheses. Common to all datasets, is the notion of inhibitive or regular behaviour at small spatial scales (5.88-$13.28\mu m$) and can be thought of as a proxy for the 2D representation of 3D cell size, which we believe to be approximately $13.18\mu m$ (s.d. 2.16). An alternative suggestion is the idea that lactotroph cells may have a preference to be located next to cells of a different type rather than next to other lactotrophs.

The additional hybrid components of each fitted model start to highlight changes as pituitary tissue develops. Specifically, we see that at embryonic day 18.5, there

Table 6.2: Estimated parameter values of the fitted hybrid Gibbs models for each dataset. Nuisance parameters (the interaction distances) were obtained by maximising the profile pseudo-likelihood and all other parameters were obtained by maximising the pseudo-likelihood. Note that the R package spatstat provides estimates of the log parameters and consequently the standard deviations are presented only on the log scale in brackets.

Dataset	Hardcore Component	Component 2	Component 3	Component 4
A1	$\hat{r}_0 = 2.62$	Strauss: $\hat{\gamma} = 2.88$ $\log \hat{\gamma} = 1.06\,(0.25)$ $\hat{r}_1 = 3.83$	Geyer-Saturation: $\hat{\gamma} = 0.80$ $\log \hat{\gamma} = -0.22\,(0.08)$ $q = 2$ $\hat{r}_2 = 7.88$	Area-Interaction: $\hat{\eta} = 3.05$ $\log \hat{\eta} = 1.11\,(1.20)$ $\hat{r}_3 = 18.27 \times 2$
A2	$\hat{r}_0 = 4.38$	Strauss: $\hat{\gamma} = 0.43$ $\log \hat{\gamma} = -0.84\,(0.48)$ $\hat{r}_1 = 5.88$	Area-Interaction: $\hat{\eta} = 1.17$ $\log \hat{\eta} = 0.17\,(0.38)$ $\hat{r}_2 = 11.65 \times 2$	–
A3	$\hat{r}_0 = 2.48$	Strauss: $\hat{\gamma} = 0.39$ $\log \hat{\gamma} = -0.94\,(0.34)$ $\hat{r}_1 = 8.27$	Geyer-Saturation: $\hat{\gamma} = 1.70$ $\log \hat{\gamma} = 0.53\,(0.21)$ $q = 4$ $\hat{r}_2 = 32.29$	–
P1	$\hat{r}_0 = 2.48$	Strauss: $\hat{\gamma} = 0.52$ $\log \hat{\gamma} = -0.65\,(0.21)$ $\hat{r}_1 = 9.94$	Area-Interaction: $\hat{\eta} = 0.33$ $\log \hat{\eta} = -1.11\,(0.96)$ $\hat{r}_2 = 16.22 \times 2$	–
P2	$\hat{r}_0 = 4.82$	Strauss: $\hat{\gamma} = 0.40$ $\log \hat{\gamma} = -0.92\,(0.39)$ $\hat{r}_1 = 13.09$	Area-Interaction: $\hat{\eta} = 2.19$ $\log \hat{\eta} = 0.78\,(0.95)$ $\hat{r}_2 = 9.09 \times 2$	–
E1	$\hat{r}_0 = 5.61$	Strauss: $\hat{\gamma} = 0.49$ $\log \hat{\gamma} = -0.72\,(0.24)$ $\hat{r}_1 = 13.28$	–	–

are no signs of further spatial dependence over an interaction distance of $13.28\mu m$, indicating cell location can be modelled as a random process conditional on cells being further than $13.28\mu m$ apart. However, as tissues develop more structure becomes apparent, with cell clustering evident in all three adult datasets up to an interaction distance of 23-36μm. It is unclear if this is also true at post-natal day 1.5 since datasets P1-P2 show differing results. Biologically, this suggests an initial random placement of cells throughout the pituitary in early development with clusters of cells appearing later on. This is perhaps unsurprising when one considers the impact of cell differentiation.

In light of the above, we make the significant caveat that these results are based on a very small sample size (six datasets in total). Consequently, the results should be viewed as hypotheses rather than conclusions.

6.4 Discussion

This section has applied spatial point process analysis to the location of lactotroph cells in tissues at different stages of development. The exploratory analysis of this chapter has shown how different networks can be constructed across the tissue to form very different levels of tissue connectivity ranging from large numbers of small clusters to small numbers of large clusters. We shall see how this perspective can be used in the spatio-temporal analysis presented in Part III.

The modelling framework presented in this chapter enables us to construct possible hypotheses regarding the connectivity of pituitary tissue in different stages of development. Specifically, we see evidence of an increase in cell connectivity as the tissue matures, as the range of interaction for spatial correlation increases. However, one should note that although hybrid models provide a consistent analysis, the fits produced from a simple hardcore Strauss process appeared reasonable for all datasets in terms of the residual summary statistics. Thus, due to the small sample size, some care should be taken when considering the impact of these results.

In the final part of this thesis, we will investigate how the spatial structure interacts with any temporal correlation. For example, we will consider both the network analysis of Section 6.2.2 and the modelling analysis of Section 6.3.2 in order to link the spatial and temporal relationships.

Part III

Spatio-Temporal Coupling of Gene Transcription

CHAPTER 7

SPATIAL TRANSCRIPTIONAL DYNAMICS

Model building is the art of selecting those aspects of a process that are relevant to the question being asked.

J.H. Holland

7.1 Introduction

This third and final part of the thesis extends the analysis of temporal single cell gene expression data into a spatial domain with a view of incorporating the spatial organisational structure as described in Chapter 6. Very little is known about the mechanisms driving Prolactin gene expression and in particular whether there exists a signalling mechanism to coordinate individual cells. Previous studies have seen trends of changing expression within intact tissue (Harper et al., 2010) with higher levels of expression found near the periphery of the tissue. However, until recently, there has been little investigation into cell signalling mechanisms within intact tissue. From the start of this project there has been the motivation to investigate different mechanisms of signalling since it is evident from the raw light intensity data that some spatial coordination exists. This is shown in Figure 7.1, which shows how the Pearson correlation coefficient of any two time series changes with the pairwise Euclidean distance (see also Appendix Figure C.7). It can be seen that in all adult datasets, and to a lesser extent in the E18.5 datasets, there is a decrease in the pairwise correlation as the distance increases. Moreover, the distance at which

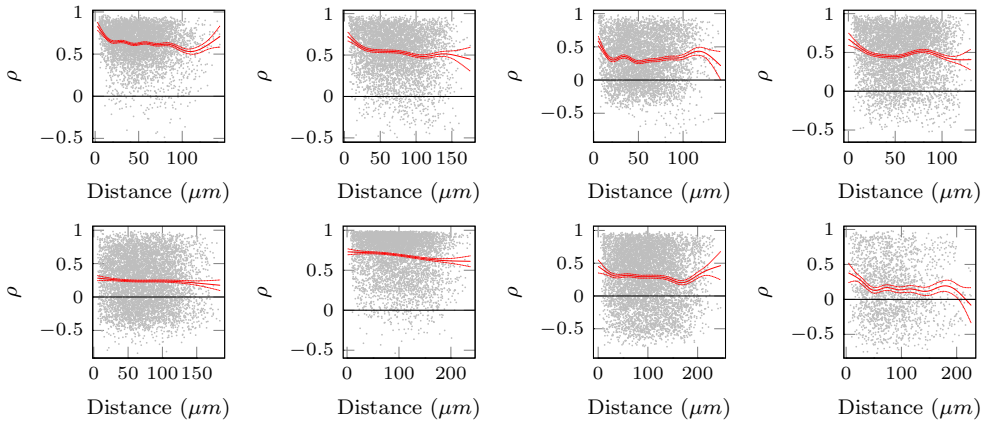


Figure 7.1: The Pearson correlation coefficient, ρ, between any two time series is plotted against pairwise Euclidean distance. The red line is the penalised regression spline with dashed lines indicating the 95% confidence bands about the mean response. The black line indicates the value of zero correlation. Top row corresponds to datasets A1-A4 and bottom row to P1-P2 and E1-E2. See also Appendix Figure C.7 for a transformed regression.

the relationship dissipates is similar to the clustering distances estimated in the hybrid Gibbs models of the previous chapter. Unfortunately, the pairwise correlation coefficient of the light intensity time series gives little insight into possible signalling mechanisms either mathematically or biologically. In this chapter, we aim to work with the back-calculated transcriptional profiles to gain more insightful analyses.

The remainder of this chapter is structured as follows, we first give an overview in Section 7.2 of the different categories of biological cell signalling that we hope to distinguish between and also the types of signalling that are not possible to extract given the data available. Section 7.3 describes the analysis of the back-calculated transcriptional profiles over space through the use of different score functions. The analysis presented in Sections 7.2-7.3 is joint work with Hiroshi Momiji, a postdoctoral researcher in Warwick Systems Biology. These analyses reveal certain spatial dependencies that have been incorporated into a mathematical framework in Section 7.4 with a simulation study presented in Sections 7.5 and 7.6. We end the chapter with a brief discussion of the key points and directions for future work.

7.2 Biological Coupling

There are many detailed biological mechanisms for cell signalling models, but given the data available, we consider only relatively coarse-grained models for our analyses, which are described below.

1. Hub Models. Hubs of coordinated activity may arise from networks of interconnected cells. For example, as depicted in Figure 6.6 of the previous chapter, a network may be constructed such that it results in clusters of disconnected cell groups, where each cluster has a coordinated response. This type of network could arise, for instance, in response to external signals, where only certain portions of the tissue respond.

2. Juxtacrine signalling. Juxtacrine signalling is a method of cell signalling through direct cell-cell contact of neighbours. Examples include the calcium signalling of the somatotroph network, which is propagated through adherens junctions (cell-cell contacts) to travel across the tissue (Bonnefont et al., 2005).

3. Paracrine signalling. Paracrine signalling occurs when cells release paracrine factors, which diffuse short distances to then induce changes in nearby cells. This signalling method does not rely on direct cell-cell contact but is still viewed as short range signalling.

There, of course, exist many other types of signalling mechanisms, for example heterotypic networks, where signals may reach lactotroph cells through interactions of another cell type. Similarly, signals could travel through the vasculature and blood capillaries between lactotroph cells. However, since we have data on neither the vasculature structure nor other cell types it is not currently possible to identify these signalling mechanisms.

Although, it is not known which signalling mechanisms may be relevant for lactotroph networks nor how the signals themselves induce transcription, one can try to identify properties of the signalling mechanism through the properties of the spatial dependence of the back-calculated transcriptional profiles.

7.3 Spatial Score Functions

Not only are there a number of possible mechanisms for spatial coupling but also a number of different targets. Since we are primarily interested in transcription, cell signalling could affect either the timings of transcriptional switches, the rate of transcription or both. In order to identify these different targets, we have utilised several score functions.

1. Pearson Correlation Coefficient, which will indicate synchronicity in transcriptional profiles based on their switch times and switch heights. Note, that only

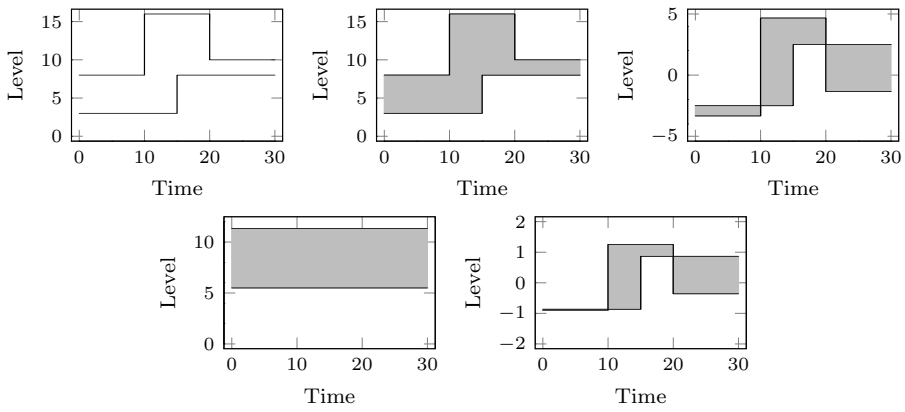


Figure 7.2: Illustration of the different spatial score functions. The shaded blue region indicates the score value for the pair of transcriptional profiles shown in the far left hand plot. From left to right the Scores shown correspond to Scores 2-4 described in the main text.

profiles with at least one switch can be used, since profiles with zero switches have ill-defined correlation.

2. Area between profiles. This measure will indicate synchronicity in transcriptional profiles based on both the timings of switches and the levels of transcription.

3. Area between mean corrected profiles. Synchronicity evaluated with this measure is based on the similarity of switch times and switch heights but disregards the relative transcriptional levels.

4. Area between mean levels. This measure will be predominantly driven by similarity in the overall transcriptional level and not in the switch times.

5. Area between "normalised" profiles. Profiles are mean corrected and if a profile has one or more switches it is normalised to have variance one.

The advantage of the area based measures defined by Scores 2-5 over the correlation coefficient of Score 1, is that the associated scores are well defined regardless of the number of switches.

Figure 7.2 illustrates Score functions 2-5 for a pair of example transcriptional profiles. These different Score functions capture differing features of the synchrony between transcriptional profiles. This can be seen in Figure 7.3, which shows how the different Scores calculated on the marginal transcriptional profiles of dataset

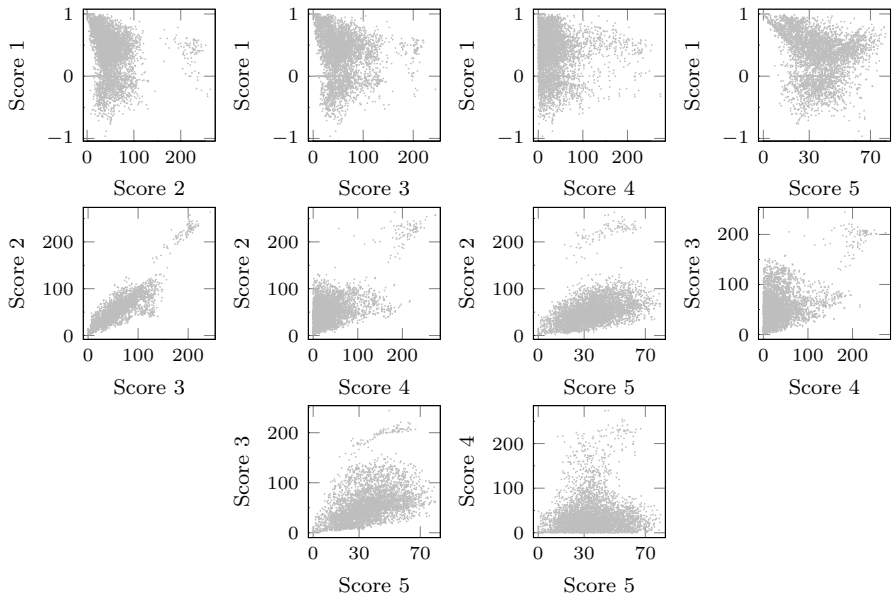


Figure 7.3: Relationship between each Score function calculated for dataset A1.

A1 relate to each other. Although there is reasonable agreement, there is still considerable deviation from a direct linear relationship. For instance, Scores 2 and 3 are highly related but there is a large amount of heteroscedsticity between Scores 1 and 3. Note that to avoid biasing the area-based measures of synchronicity, we require an appropriate weight related to the length of the overlap between time series. In addition, note that increased synchronicity is given by smaller scores for Score functions 2-5 and by larger scores for Score function 1.

To identify the possible biological mechanisms of coupling we investigate the relationship of these Score functions over pairwise distance. In particular, evidence of hub cells would be indicated in clustering of similar cells across space, whereas evidence of juxtacrine or paracrine signalling would be indicated through a dependence between the Score function and the pairwise distance.

Recall, that there are several ways in which one can summarise the posterior transcriptional profile, namely the marginal transcriptional profile obtained by averaging over the profiles of all the MCMC iterations or a conditional approach, which extracts all possible transcriptional profiles with an associated probability of occurrence for each cell. Although, we have considered both the spatial relationship of the posterior marginal profiles and the weighted conditional profiles, the analysis presented in this section shows only the marginal approach. This is because, typically, the weighted conditional profiles contain too few switches to accurately

calculate the associated scores. Consequently, the interpretation of synchronicity of the marginal profiles is that the average cell behaviour is synchronised rather than any single realisation.

Distance Based Analysis

Considering the five different Score functions, we found evidence of spatial synchronisation only under Scores 1, 2, and 4. Scores 3 and 5 revealed little dependence on pairwise distance in all datasets both under the marginal transcriptional profiles and the weighted average of the conditional transcriptional profiles. This suggests that synchrony is not exclusively due to the switch times in transcription but also to the change in height associated with each switch.

Score 2 revealed some spatial dependence, however, this dependence was not consistent across all datasets. This can be seen in Figure 7.4, which shows the relationship between Score 2 and pairwise distance for datasets A1-A4. In particular, datasets A1 and A3 indicate a long range relationship. In contrast, A2 and A4 indicate no spatial relationship with Score 2. The behaviour of Score 4 was remarkably similar to Score 2.

The most consistent behaviour was found in Score 1. Figure 7.5 shows the behaviour of Score 1 over pairwise distance, which shows similar patterns to the spatial relationship of pairwise correlation of the raw time series shown in Figure 7.1. In particular, there is increased synchronicity at short ranges in datasets A1-A3 that is not evident in the four immature tissues. Moreover, if instead the score is calculated on the log transcriptional profiles as shown in Figure 7.6, there is evidence of short range synchronicity in all adult datasets.

Consequently, this distance based analysis indicates a short-range synchronicity of the transcriptional switch times in the adult tissue and perhaps a limited mid-range synchronicity of transcriptional levels also in the adult tissue. In contrast, there is no evidence of synchronised transcriptional profiles of either the timings or the levels for the immature tissues.

Cluster Based Analysis

An alternative to the distance based analysis, is to look for clusters of similarly behaved cells. This can be achieved by clustering based on specific cell features. For instance, we consider the following four features.

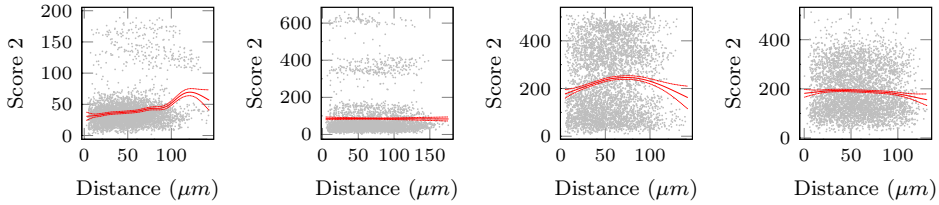


Figure 7.4: The relationship between Score 2 (calculated on the log transcriptional profile) and Euclidean distance for datasets A1-A4. A Box-Cox transformation has been applied to Score 2 to allow a penalised regression spline to be fitted (shown in red). Thus, the y-scales are not comparable between datasets.

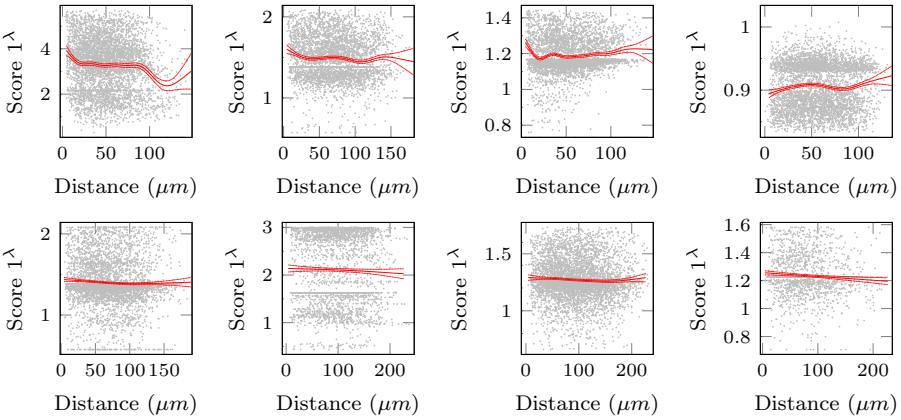


Figure 7.5: The relationship between Score 1 and Euclidean distance for datasets A1-A4 (top row) and P1-P2,E1-E2 (bottom row). A Box-Cox transformation has been applied to Score 1 to allow a penalised regression spline to be fitted (shown in red). Thus, the y-scales are not comparable between datasets.

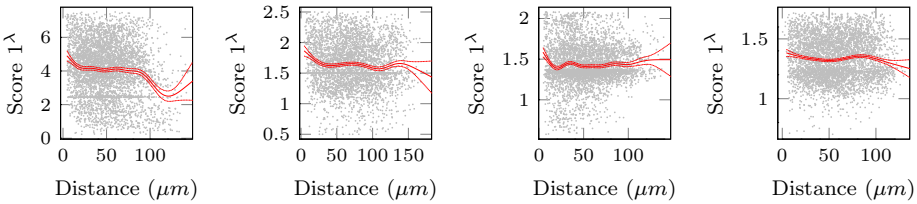


Figure 7.6: The relationship between Score 1 (calculated on the log transcriptional profile) and Euclidean distance for datasets A1-A4. A Box-Cox transformation has been applied to Score 1 to allow a penalised regression spline to be fitted (shown in red). Thus, the y-scales are not comparable between datasets.

1. Feature 1. The median (over time) transcriptional rate. This feature will be indicative of the average transcription rate for an individual cell.

2. Feature 2. The mean transcriptional rate, weighted by the time spent in each transcriptional state. As with Feature 1, this will be indicative of the average transcriptional state of a cell.

3. Feature 3. The dynamic range of transcription, calculated as the maximum rate minus the minimum (over time). This will represent how dynamic each cell is in terms of their transcriptional behaviour.

4. Feature 4. The total number of switches in a transcriptional profile. Again, this will represent how dynamic a cell is in the transcriptional behaviour, based upon the number of switches rather than the range of levels.

These spatial features can be viewed as a marked spatial point process, extending the analysis presented in Part II. Specifically, each point of a process has associated a "mark", which in this example will be a specific transcriptional feature. Clustering of discrete marks can be analysed through marked summary statistics, for example the multivariate K-function. However, the features we investigate are, in general, continuous and it is difficult to quantitively assess spatial clustering or dependence. One approach is to split the data into bins and view the process as a discretely marked point process. The output of such a method is shown in Figures 7.7-7.10 for the various spatial Features binned into 10 quantiles. However, analysing this as a discretely marked point process is highly sensitive to the definition of the bins and moreover, the dependence between bins is not taken into account. In order to provide some quantitative assessment of spatial dependence of the marks, we instead calculate the associated variograms given in Appendix Figures C.8-C.11 for each of the four spatial features. A brief description of the variogram is given in Appendix B.3. It should be noted that the variogram is defined for a continuous spatial process rather than a point process and as such may not fully represent the data we have here.

In general, the qualitative spatial behaviour of these features when calculated on the marginal transcriptional profiles compared to the behaviour calculated based on the weighted conditional transcriptional profiles was highly consistent. Figures 7.7-7.10 show the spatial distribution of these features calculated from the marginal transcriptional profiles for Features 1-4 respectively. Unsurprisingly, the spatial behaviour of the median and weighted mean transcription rates is similar and, moreover, shows distinct clustering, particularly in datasets A1-A3, where similarly be-

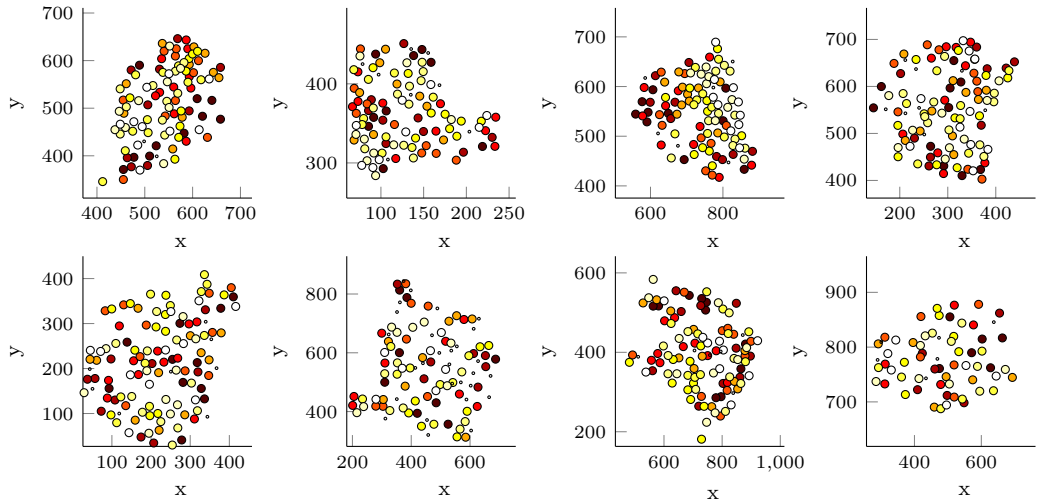


Figure 7.7: The spatial distribution of Feature 1 for datasets A1-A4 (top row) and P1-P2, E1 and E2 (bottom row). Colour intensity corresponds to the sorted value of Feature 1 over all the cells, with each colour representing a decile of the data. Black corresponds to the lowest decile and white to the highest decile. Small dots indicate cells with a missing value.

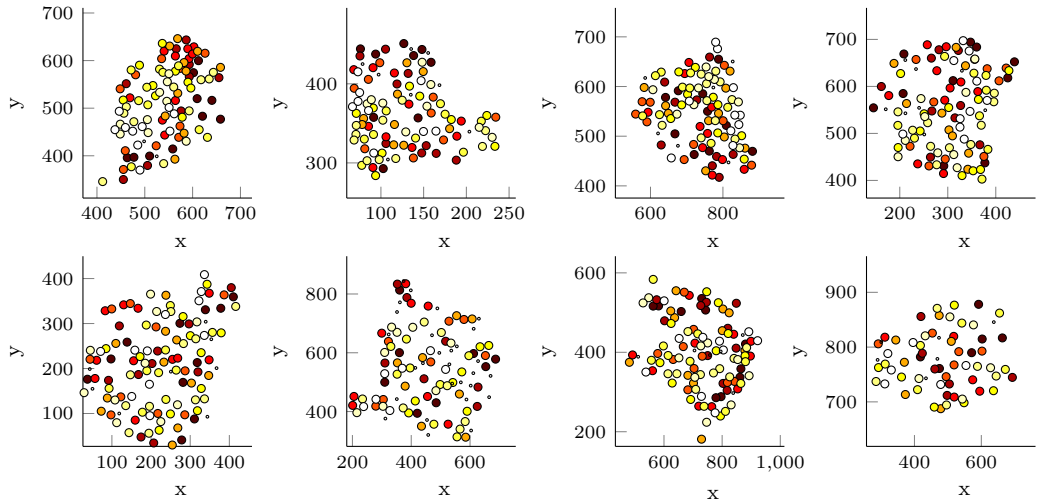


Figure 7.8: The spatial distribution of Feature 2 for datasets A1-A4 (top row) and P1-P2, E1 and E2 (bottom row). Colour intensity corresponds to the sorted value of Feature 1 over all the cells, with each colour representing a decile of the data. Black corresponds to the lowest decile and white to the highest decile. Small dots indicate cells with a missing value.

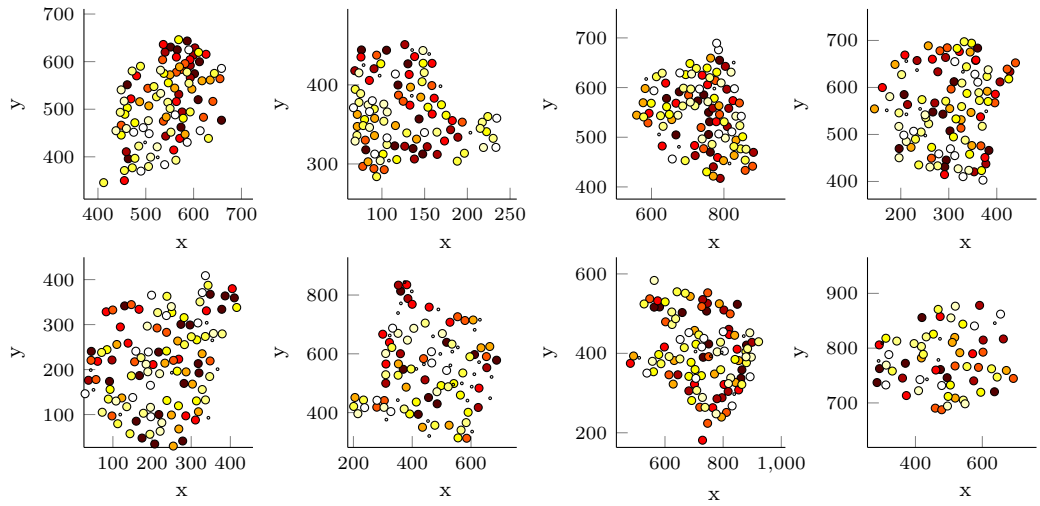


Figure 7.9: The spatial distribution of Feature 3 for datasets A1-A4 (top row) and P1-P2, E1 and E2 (bottom row). Colour intensity corresponds to the sorted value of Feature 1 over all the cells, with each colour representing a decile of the data. Black corresponds to the lowest decile and white to the highest decile. Small dots indicate cells with a missing value.

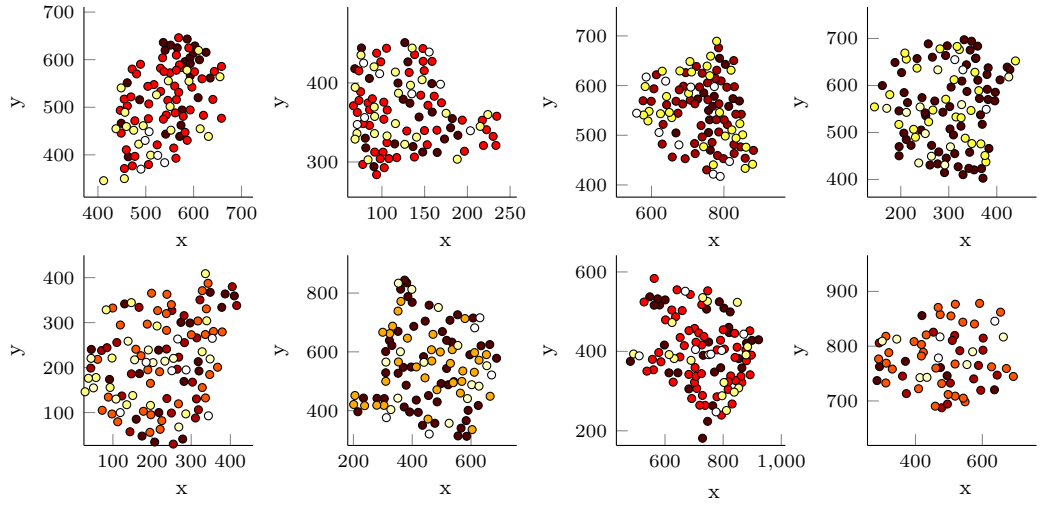


Figure 7.10: The spatial distribution of Feature 4 for datasets A1-A4 (top row) and P1-P2, E1 and E2 (bottom row). Colour intensity corresponds to the sorted value of Feature 1 over all the cells, with each colour representing a decile of the data. Black corresponds to the lowest decile and white to the highest decile. Small dots indicate cells with a missing value.

haved cells are all located together (as supported by the variograms in Appendix Figure C.8). It appears, that the immature tissues show evidence of a tighter coupling in the sense that similar features are often seen in direct neighbours but not in large clusters. This is difficult to validate quantitively since the variograms cannot clearly detect similarity in direct contacts. Evidence of clustering is further reduced in Features 3 and 4 for all datasets.

Consequently, this cluster based analysis indicates evidence of a networked synchronisation of the overall transcriptional rate, where clusters of similarly behaved cells are located together. In addition, there is some evidence of cells being coupled into small groups based on their dynamic range for the adult tissues that is not found in the immature tissues.

In light of the above analysis, we look at investigating the following two mechanisms for spatial coupling.

1. Overall levels of transcription may be regulated in hubs or portions of tissue, as measured in the clustering of Features 1 and 2.

2. Synchronicity of switch times and switch heights (as measured through the Pearson correlation coefficient or Score 1) is regulated across short range distances, in mature tissue.

Moreover, the above mechanisms are evident in most adult datasets but little evidence exists for synchronicity in the younger tissues and if there is synchronicity in the transcriptional levels, this occurs in much smaller tighter clusters in the immature tissue compared to the adult. We therefore believe that spatial synchronicity and coordination develops as the tissue matures as supported by Featherstone et al. (2011).

It is ongoing work within our research group to elicit further details of the biological mechanisms that can explain the observed synchronicity. In particular, there is ongoing investigation attempting to distinguish between juxtacrine and paracrine signalling mechanisms through the incorporation of the cell size. Coupled with this, we have also investigated how the spatial dependence can be incorporated into the transcriptional profiles and is considered in detail in the following section.

7.4 Spatial Likelihood

The aim of this section is to extend the models for transcription, derived in Chapter 4, to incorporate a spatial structure. For now, we restrict ourselves to formulating models of spatial coordination based on the evidence shown in adult tissues, namely hubs of similar levels and synchronised switches over short range distances. To do this, recall from Section 4.5.1 the parametric model for transcription defined through the transition probabilities given by,

$$f(\beta^{(i)}|\theta) = \prod_{j=0}^{K^{(i)}} \mathbb{P}(t_{j+1}^{(i)} - t_j^{(i)}, \beta_{t_{j+1}}^{(i)} | \beta_{t_j}^{(i)}), \tag{7.1}$$

where $t_{j+1}^{(i)} - t_j^{(i)}$ are the inter-switch times of cell i and $\beta^{(i)}$ is the associated transcriptional profile, satisfying,

$$\beta^{(i)}(t) = \beta_{t_j} \mathbb{I}_{[t \in (t_j, t_{j+1})]}, \quad \text{for } j = 1, ..., K^{(i)}.$$

Here $K^{(i)}$ is the number of switches with $t_1^{(i)}, ..., t_K^{(i)}$, the timings of each transcriptional switch for cell i and $t_0^{(i)} := 0$. Moreover, letting $\Delta_j := t_{j+1} - t_j$, we previously derived,

$$\mathbb{P}(\Delta_j, \beta_{t_{j+1}}) | \beta_{t_j}) = q(\Delta_j)\, \omega(\log \beta_{t_{j+1}} | \log \beta_{t_j}),$$

where q and ω are given by,

$$q \sim Exp(\lambda_1) + Exp(\lambda_2), \tag{7.2}$$

$$\omega \sim \mathbb{P}(\text{up}|\beta(t)) N(\mu_{\text{up}}(t), \sigma^2) + \mathbb{P}(\text{down}|\beta(t)) N(\mu_{\text{down}}(t), (\sigma/a_1)^2), \tag{7.3}$$

such that,

$$\mu_{\text{up}}(t) = a_0 + a_1 \log \beta(t),$$
$$\mu_{\text{down}}(t) = (a_0 - \log \beta(t))/a_1.$$

Thus, the time to next switch is assumed to follow a sum of two Exponential distributions and the relationship between any consecutive rates follows a log-linear relationship. Referring back to Section 4.5.1, these properties were derived from the posterior transcriptional profiles. The associated biological interpretation of this model is that cells switch at random between different transcriptional states condi-

tional on the cell first switching to a refractory or recovery state before moving to a new transcriptional state. In addition, the relationship between consecutive rates, although only weakly informative, can be modelled through a log-linear relationship. Thus, to keep the same biological interpretation, we look at extending the derived temporal transcriptional model into the spatial domain.

In its most general form, the spatial extension to the likelihood given in equation (7.1) is defined to be,

$$f(\beta^{(1)},...,\beta^{(N)}|\theta) = \prod_{s=1}^{N}\left[\prod_{j=1}^{K^{(s)}}\mathbb{P}\left(t_{j+1}^{(s)}-t_{j}^{(s)},\beta_{t_{j+1}}^{(s)}|\beta_{t_j}^{(s)},\beta^{(1)},...,\beta^{(s-1)},\beta^{(s+1)},...,\beta^{(N)}\right)\right], \tag{7.4}$$

where the transition probability is allowed to depend upon the transcriptional profile of neighbouring cells. Clearly, this dependence on neighbouring profiles can take many forms but in line with equations (7.2) and (7.3), we restrict the spatial extensions to take the following form,

$$q \sim Exp(\lambda_1) + Exp(\lambda_2(s)), \tag{7.5}$$

$$\omega \sim \mathbb{P}(\text{up}|\beta(t))N(\mu_{\text{up}}(s,t),\sigma^2) + \mathbb{P}(\text{down}|\beta(t))N(\mu_{\text{down}}(s,t),(\sigma/a_1)^2), \tag{7.6}$$

where λ_2 is a function of space and $\mu(s,t) := (\mu_{\text{up}}(s,t),\mu_{\text{down}}(s,t))$ is a function of both space and time. The above spatial extension has the desirable property that under spatial independence, the temporal transcriptional model is recovered.

Spatial dependence has been introduced in two parameters. Specifically, synchronicity of switch times is incorporated through the parameter λ_2 and will be discussed in more detail in Section 7.4.1. Notice that we do not allow λ_1 to depend upon space. This can be justified since λ_1 represents a transition to a refractory state and is the inverse of the mean cell recovery time and consequently should not depend upon the activity of neighbouring cells.

Synchronicity of transcriptional levels is incorporated into the parameter μ and will capture the behaviour of neighbouring cells. No further spatial dependence is currently given to the parameter σ^2 for simplicity.

There are many different forms of the spatial dependence one can impose on the two spatial parameters. In particular, if one knew the mechanistic features such as whether juxtacrine or paracrine signalling is present, one could construct a paramet-

ric, biologically interpretable model. However, since this is not known, we provide a brief discussion of several semi-parametric forms of the spatial dependence that may be used to test such mechanistic hypotheses.

7.4.1 Spatial Transition Times

Firstly, consider the transition parameter λ_2. This represents the rate of transitions to a new transcriptional state out of a refractory period. As seen in the previous section, there's evidence of increased synchronicity of switch times for cells that are located close together. Consequently, a reasonable form for λ_2 is given by,

$$\frac{1}{\lambda_2(s)} = \begin{cases} \frac{1}{\lambda_{\max}} & \text{if no neighbours switch} \\ \frac{1}{\lambda_{\max}} p(s) & \text{otherwise,} \end{cases} \tag{7.7}$$

where $p(s) < 1$ such that when a neighbour switches, the mean transition time decreases. Let $\mathcal{N}$ denote the set of neighbours within a distance D and let $\mathcal{N}_s \subset \mathcal{N}$ be the subset that have switched in the last T time. We consider three candidate models for $p(s)$ given below,

1. $p_1(s) = \prod_{c \in \mathcal{N}_s} K \left(\frac{dist(s,c)}{D} \right)$.
2. $p_2(s) = \prod_{c \in \mathcal{N}_s} K \exp \left(\frac{dist(s,c) - D}{D} \right)$.
3. $p_3(s) = \prod_{c \in \mathcal{N}_s} K \frac{dist(s,c)}{D} \exp \left(\frac{dist(s,c) - D}{D} \right)$.

All three models have the property that all neighbours of a cell influence the transcriptional dynamics independently with identical structure. In particular, model 1, has a linear dependence with pairwise distance. In contrast both models 2 and 3 have an exponential dependence on pairwise distance and are reminiscent of exponential models for spatial random fields. In a similar way, one can envisage constructing further models of spatial dependence analogously to the Gaussian, spherical or Matérn models for spatial random fields (Diggle and Ribeiro, 2007). At present, we restrict attention to the three models prescribed above.

In each of the models, the parameter K controls the strength of synchronicity, i.e. the higher K is, the less influence the neighbouring cells have. Figure 7.11 shows how the different definitions of p behave as a function of distance to a neighbouring cell for $K = 1$ and $D = 10$. In particular, p_1 and p_2 both have the property that at a distance $dist = 0$, $p = 0$. Note that in the above definitions, we assume $dist(s, c)$

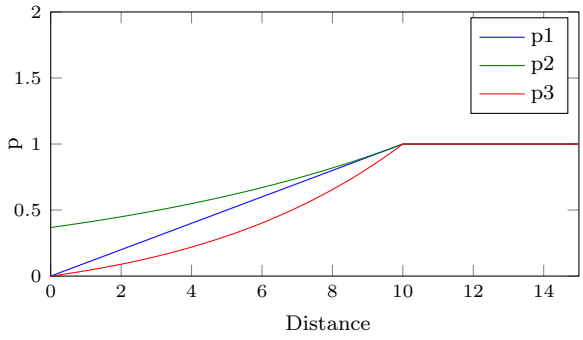


Figure 7.11: Illustration of how the parameter p varies over distance under three different definitions given in the main text.

to be the Euclidean distance between cells c and s. All three definitions of p assume an instantaneous signal that does not decay over time, i.e. all neighbouring cells that have switched in the last T time have equal influence. In order to incorporate a "decaying" signal, letting $t(c)$ denote the time since cell c last switched, we can define two further models.

4. $p_4(s) = \prod_{c \in \mathcal{N}_s} K \exp\left(\frac{dist(s,c)t(c) - DT}{DT}\right)$.

5. $p_5(s) = \prod_{c \in \mathcal{N}_s} K \frac{dist(s,c)}{D} \frac{t(c)}{T} \exp\left(\frac{dist(s,c)t(c) - DT}{DT}\right)$.

These semi-parametric models can be linked to mechanistic features. For example, if signalling was purely juxtacrine (direct cell-cell contact) the distance D would be equal to the cell size. Moreover, if the signalling was indirect, for example if a signal initiates a cascade of reactions before affecting transcription of a neighbouring cell, one might expect T to be large and preference to be given to model 5.

7.4.2 Spatial Transcriptional Levels

The parameter μ represents the mean transcriptional state of a cell after a switching event. As seen in Section 7.3, it is suspected that this parameter will depend upon spatial location, since there is evidence of spatial clustering of the median transcriptional level.

Consequently, we consider the following model for the mean log transcriptional rate μ,

$$\mu_{\text{up}}(s,t) = \nu_\beta(a_0 + a_1 \log \beta_t^{(s)}) + (1 - \nu_\beta)m(s,t),$$

$$\mu_{\text{down}}(s,t) = \nu_\beta(a_0 - \log \beta_t^{(s)})/a_1 + (1-\nu_\beta)m(s,t),$$

where a_0, a_1 are defined through the log-linear relationship of equation (7.3) and ν_β is a weight with $m(s,t)$, a location specific time varying level. Specifically, we consider $m(s,t)$ to be of the following form,

$$m(s,t) = \tfrac{1}{|\mathcal{N}_s|} = \sum_{c \in \mathcal{N}_s} \left(\log \beta_t^{(c)} \frac{w(c)}{\sum w(\nu)} \right), \tag{7.8}$$

and represents the weighted transcriptional mean of neighbouring cells. As with the definition of p in the spatial transition times, the weights could take several forms,

1. $w_1(c) = \frac{D}{dist(s,c)}$.
2. $w_2(c) = \frac{1}{\exp\left(\frac{dist(s,c)}{D} - 1\right)}$.
3. $w_3(c) = \frac{D}{dist(s,c)\exp\left(\frac{dist(s,c)}{D} - 1\right)}$.
4. $w_4(c) = \frac{1}{\exp\left(\frac{dist(s,c)t(c)}{DT} - 1\right)}$.
5. $w_5(c) = \frac{DT}{dist(s,c)t(c)\exp\left(\frac{dist(s,c)t(c)}{DT} - 1\right)}$.

In all cases, the weights increase as cells are closer to each other. In particular, the weights introduced here are analogous to the different definitions of p, describing the spatial influence of neighbouring cells on the transition times of the transcriptional profiles. Consequently, neighbouring cells affect both the rate at which cells transition to a new transcriptional state and the level at which the cells transcribe.

For the remainder of this chapter, model i will refer to the spatial transcriptional model given in equations (7.5)-(7.6) with $p = p_i$ and $w = w_i$, for $i = 1, ..., 5$.

7.5 Simulation Model

Through simulations we can investigate the behaviour of the proposed models. Specifically, we look at reproducing the spatial behaviour of the observed data where,

a) average transcriptional behaviour occurs in hubs or clusters, and
b) correlation of transcriptional profiles decreases with Euclidean distance.

7.5.1 Hub Behaviour

To reproduce the observed hub behaviour, we see the importance of taking into account the spatial organisation of cells. Specifically, the hub behaviour can be reproduced by any of the above models once cell positions are simulated from an inhibitive hardcore Strauss process. Moreover, it is difficult for the above models to exhibit the hub behaviour when cell locations are simulated from a completely random spatial process. This is shown in Figure 7.12 where transcriptional profiles are simulated from model 3, with cell locations simulated from a) a Poisson process and b) an inhibitive hardcore Strauss process. In this figure, it can be seen that although both processes reproduce the spatial correlation dependence, only the hardcore Strauss process reproduces the hub clustering behaviour, shown by the fitted variograms (see Appendix B.3 for a description of variogram behaviour).

Although one could try to incorporate the hub behaviour directly into the transcriptional model formulation, for example, defining neighbours through a network distance rather than Euclidean distance, we see no reason to do this, since the non random positioning of the cell locations is sufficient to reproduce the behaviour observed in the real data. This is exemplified in Section 6.2.2, where Euclidean networks could be constructed from cell positions to form distinct clusters of disconnected cell groups.

7.5.2 Correlation Behaviour

Since the hub behaviour is captured by the non-random spatial positioning, we now investigate how the different formulations of the spatial transcriptional model effect the spatial correlation (i.e. the relationship between pairwise correlation of the transcriptional profiles and the Euclidean distance). It is interesting to note that although we suspect the key feature driving this relationship to be the switch times, we found that one needs the spatial dependence in both the transition times and the transition rates to reproduce this behaviour. For instance, Figure 7.13 shows the correlation distance plots for transcriptional model 5 where a) the dependence is only in the transition times, b) only in the transition rates and c) in both components of the transcriptional model.

In addition to the above, although models 2 and 4 can reproduce spatial dependence, it is typically much less pronounced than in either the observed data or the other models. This is shown in Figure 7.14, where each model has been simulated with

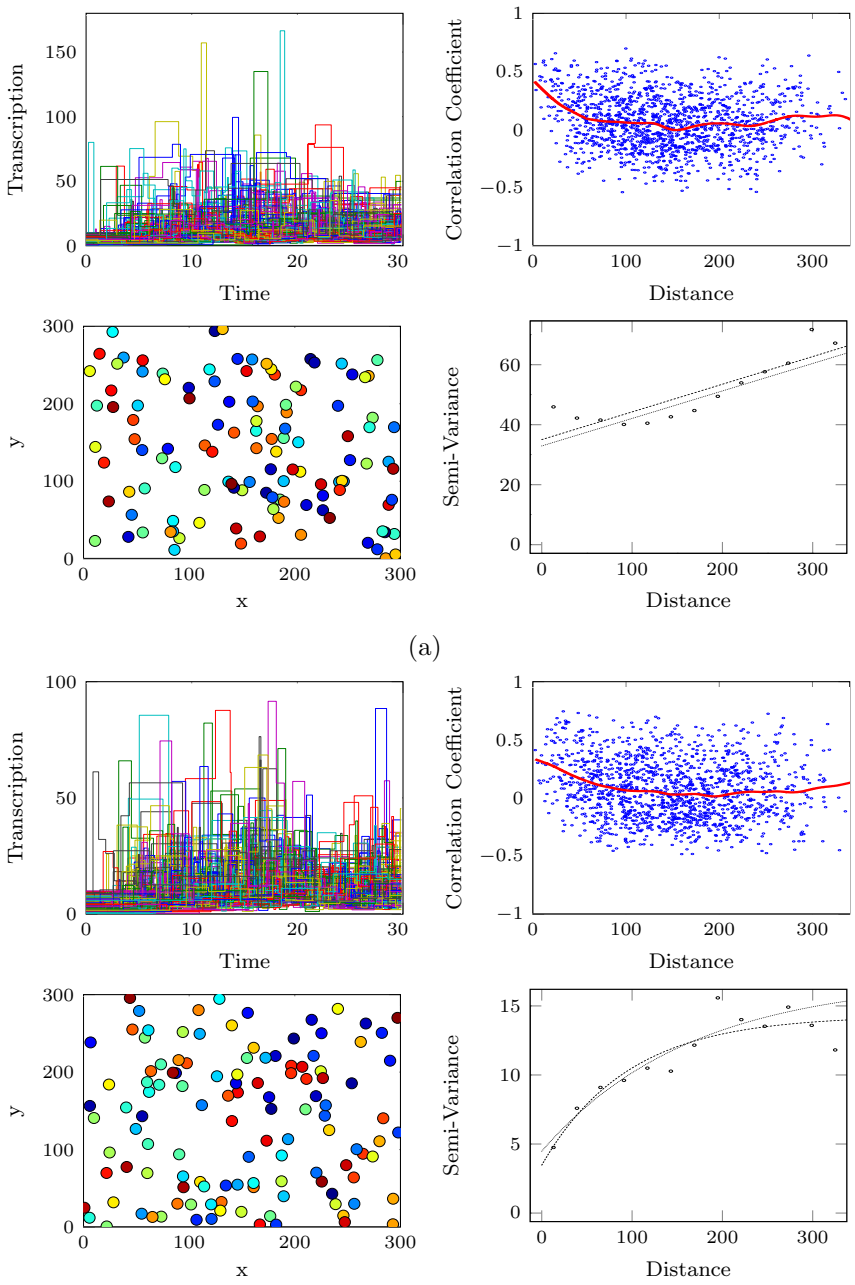


Figure 7.12: Simulating two spatial transcriptional processes from model 3 with spatial location defined via a) a Poisson Process and b) a hardcore Strauss process. The first plot in each panel shows the individual transcriptional profiles, second plot the correlation-distance relationship, the third plot the spatial location of cells with colour indicative of the mean transcription rate for each cell, weighted by the time spent in each transcriptional state (i.e. Feature 2) and the fourth plot is the associated variogram.

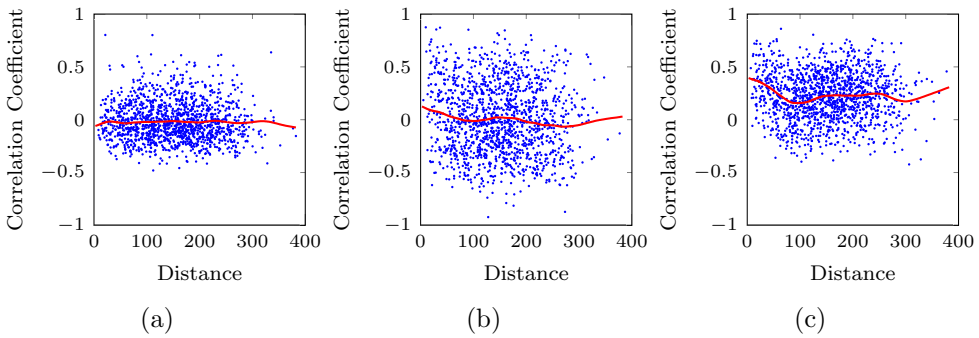


Figure 7.13: The correlation-distance relationship for transcriptional profiles simulated from model 5 where in a) spatial dependence is introduced only through the transition times, b) only through the transition rates and c) both components of the spatial transcriptional model.

the same parameter choices with the same cell location.

Thus, it seems that the definitions of p_2 and p_4 are not sharp enough (Figure 7.11) and it is desirable to have the property that at a distance $dist = 0$, p should be equal to 0. Consequently, models 1, 3 and 5 more accurately reproduce the features found in the observed adult data.

It therefore remains to investigate the effect of each of the parameters in each spatial transcriptional model. To do this systematically, we consider each parameter in turn to give a one-dimensional transect for the models 1, 3 and 5.

Threshold Distance, D

As one increases the threshold distance, the sharp decline in the correlation distance plots dissipates to the point that above a certain distance, no spatial relationship can be observed as the whole tissue is connected. As more and more of the tissue becomes connected, although the gradient of the spatial correlation decreases, the overall correlation of the tissue increases and is shown in Figure 7.15 for transcriptional model 1.

Threshold Time, T

As can be expected for a threshold time defined too small, all spatial correlation is destroyed since a cell will have no neighbours that switched in the last T time. As T increases, the spatial correlation is refined with less variability seen in the correlation distance plots, shown in Figure 7.16 for transcriptional model 3.

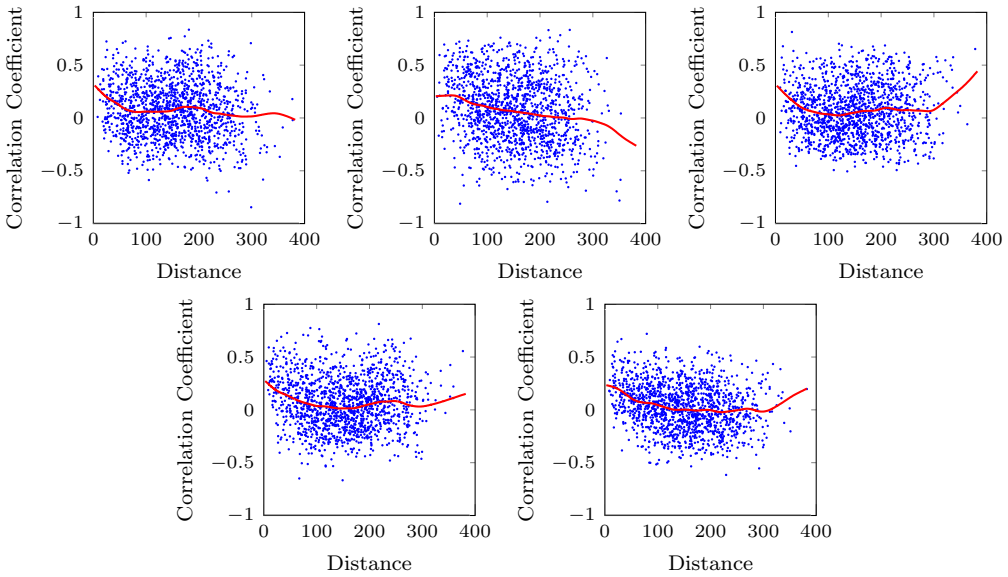


Figure 7.14: The correlation-distance relationship for each of the five possible transcriptional models simulated with the same parameter set. Left to right corresponds to transcriptional model 1-5.

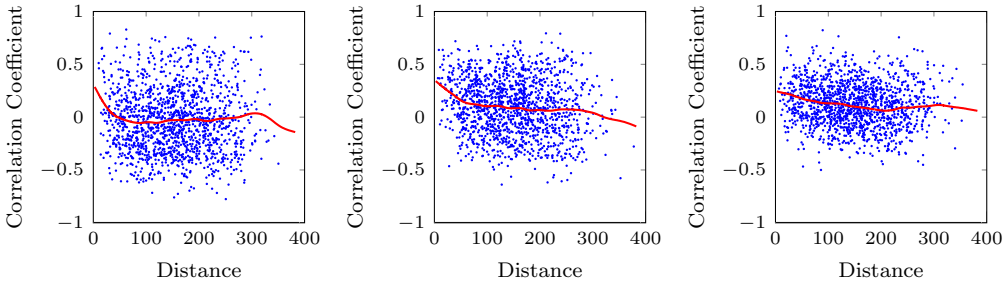


Figure 7.15: The correlation-distance relationship for possible transcriptional model 1 as the threshold distance, D increases (from left to right $D = 25, 50, 100$).

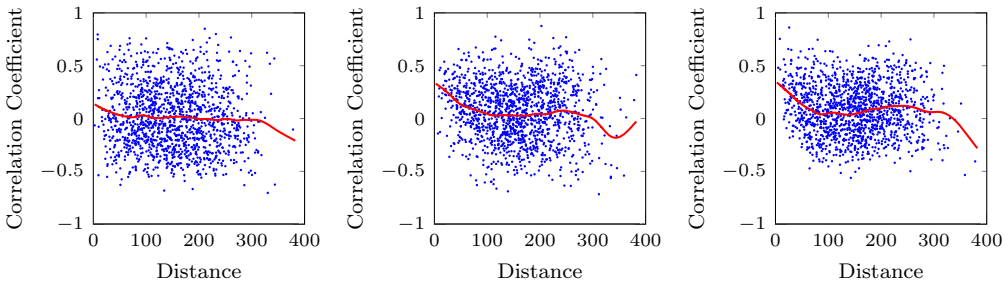


Figure 7.16: The correlation-distance relationship for possible transcriptional model 1 as the threshold time, T increases (from left to right $T = 0.25, 1, 5$).

Synchronicity Strength, K

The synchronicity parameter, K, has a similar effect as T in that as K gets smaller, the variability of the correlation coefficient between pairs of profiles decreases (plots not shown).

Spatial weight, ν_β

Finally, we consider the weight parameter, ν_β. The system is quite sensitive to the choice of this parameter with weight defined too high and no spatial correlation is observed, but with weight too low, the entire tissue can become synchronised.

Temporal Parameters

In addition to each of the above spatial parameters, it is the aim to also infer the temporal parameters, $\lambda_1, \lambda_{\max}, a_0, a_1$ and σ^2. Figure 7.17 shows transects through the likelihood under model 1. Most interestingly, the profile likelihood for parameter a_0 has two clear modes due to the mixture term in the definition of the rate transition density, ω (Equation (7.3)). This feature is common to all five transcriptional models. In addition, Figure 7.18 shows the bivariate likelihood surfaces for a single parameter set under model 1. This gives a snapshot indication of how the different parameters influence each other. In particular, there is a strong correlation between the parameters a_0 and a_1. Moreover, the threshold distance, D, is quite strongly correlated with both λ_1 and $\lambda_{\max}$, which in turn are correlated with each other.

Thus, when inference is performed, these correlation structures will be important for identifying issues of parameter identifiability and where prior information should be incorporated with most effect.

7.6 Towards an Inferential Framework

The previous section showed how three different parametric semi-mechanistic models for spatial transcription can reproduce the observed spatial relationships found in real data. The question remains that given the data available, can one distinguish between different signalling mechanisms. With a view to an inferential framework, we also wish to investigate the identifiability of these models.

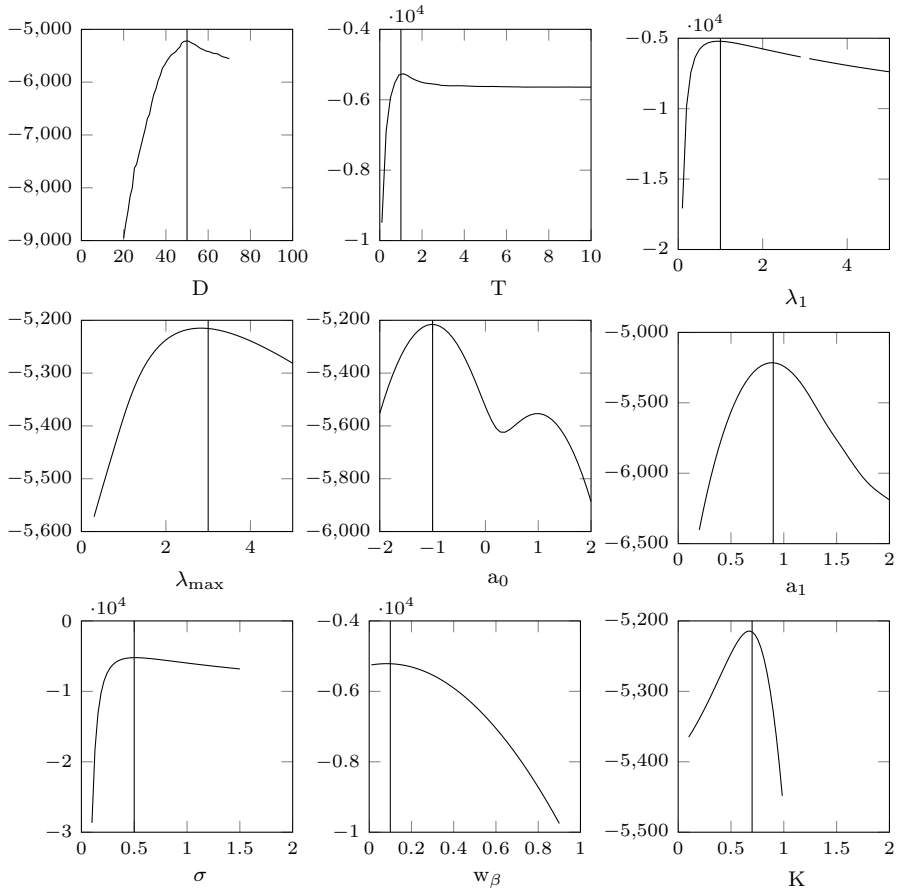


Figure 7.17: Profile Likelihood transects of spatial transcriptional model 1. True values are shown by the black lines.

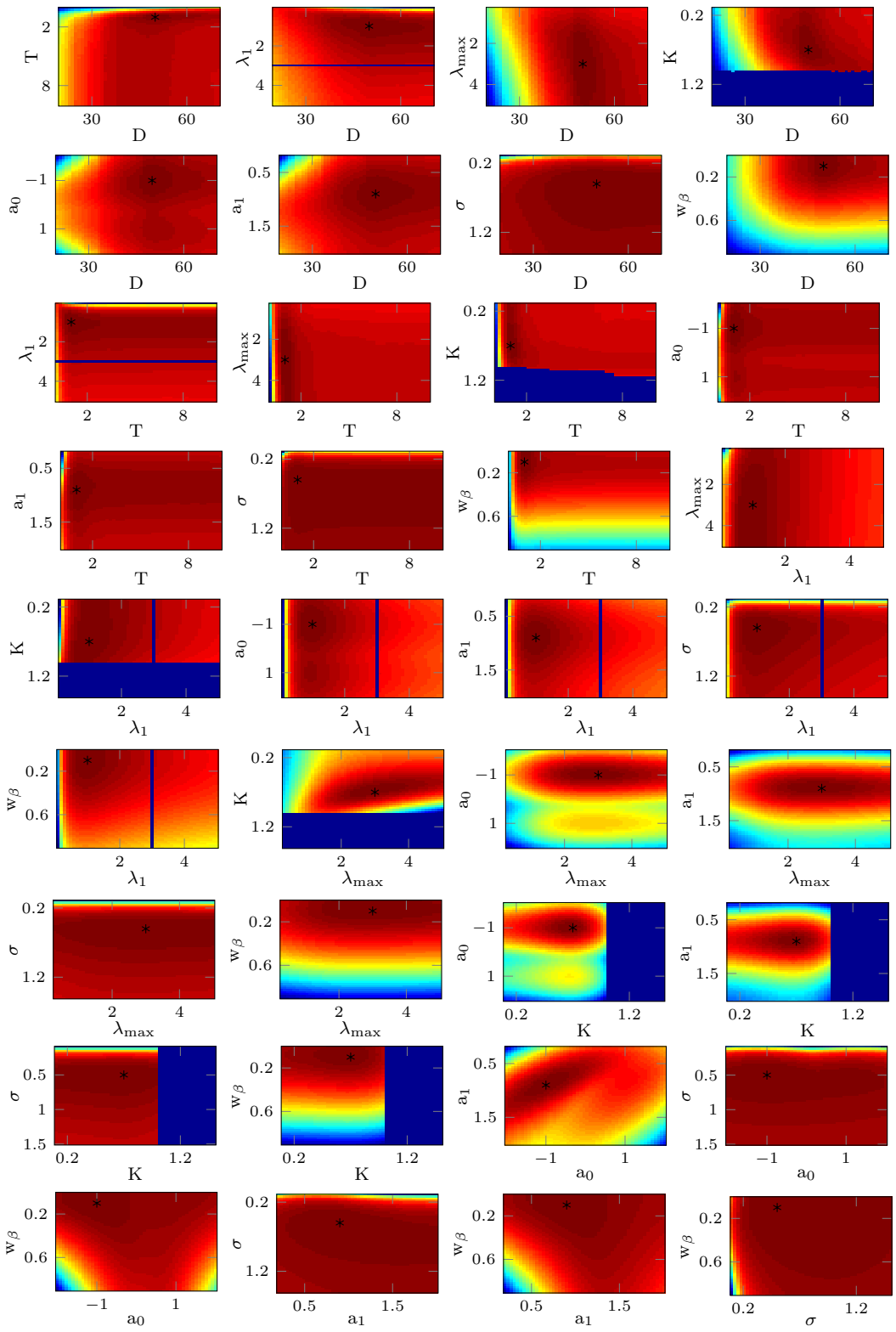


Figure 7.18: Bivariate likelihood surface of spatial transcriptional model 1. True values are shown by the black points.

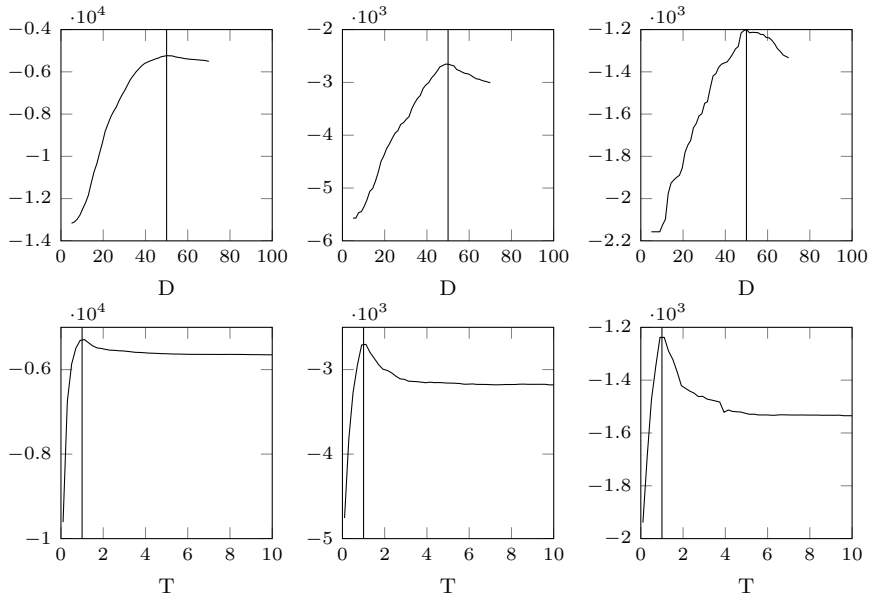


Figure 7.19: Profile likelihood transects for the threshold parameters D (top row) and T (bottom row) as the number of cells varies, from left to right $N = 119, 71, 35$. All simulations are based on transcriptional model 3.

7.6.1 Parameter Identifiability

We have already seen in Figures 7.17 and 7.18 profile and bivariate transects through the likelihood surface for a single parameter set for transcriptional model 1. From this, it is clear that the bimodality of the parameter a_0 is likely to have a large impact on inference since the bimodality affects at least three other parameters including $\lambda_{\max}$, K and a_1 (Figure 7.18). Thus, having strong prior information on a_0 will greatly aid in the identifiability of these other parameters. Alternatively, it may suggest the use of a tempered MCMC algorithm in order to efficiently explore all the possible modes of the posterior.

The other key feature associated with parameter identifiability is the resolution of the data. Here, resolution is made up of three different components.

1. The number of cells.
2. The number of switches.
3. The number of neighbours.

In order to investigate the effects of these differing data criteria, we have simulated each transcriptional model under different data scenarios restricted to the same spatial window. Firstly, we consider the profile transects through the likelihood surface to find similar effects. Namely, as the resolution of the data decreases in any

of the above criteria, the profiles of the threshold distance, D and threshold time, T, become noisier as shown in Figure 7.19 for varying numbers of cells in model 3.

It is unsurprising that the resolution components 1-3 have similar effects, since all three effectively define the number of neighbours that contribute to the spatial dependence in each model. Consequently, due to the increased noise in the likelihood, for poor data resolution, or equivalently, little spatial dependence, estimating the spatial models will become challenging.

To perform inference on the three different transcriptional models (models 1, 3, and 5 of Section 7.4) we consider four different data scenarios:

1. Fast switching dynamics and highly connected tissue. This is characterised by an average of 7.5 switches in a time period of 30 hours with the average number of connected neighbouring cells approximately 19.

2. Fast switching dynamics and sparsely connected tissue. A sparse tissue was simulated so that the average number of neighbours was approximately 2.5.

3. Slow switching dynamics and highly connected tissue. Slow switching dynamics were characterised by an average of 3.75 switches in a time period of 30 hours.

4. Slow switching dynamics and sparsely connected tissue.

For each scenario, we ran a simple Metropolis-Hastings random walk MCMC algorithm to target the posterior distribution. In all cases, we assumed uninformative prior distributions over each of the different model parameters. Performing inference on each of the models in various different data scenarios, we found the posterior to be well identified about the true value so long as the switch dynamics are strong enough. Explicitly, we found that if the parameter a_1, which defines the size of transcriptional switches, is too small, estimation becomes difficult, and the Markov chains have a tendency to explore the wrong mode in the posterior. However, for values of a_1 similar to that observed in real data (Table 4.3), we obtain reasonable posterior estimates as shown in Figure 7.20 for transcriptional model 3 assuming a sparsely connected network of cells.

7.6.2 Model Identifiability

Having investigated the parameter identifiability within each individual model it is also of interest to know if one can distinguish between the different models. To do

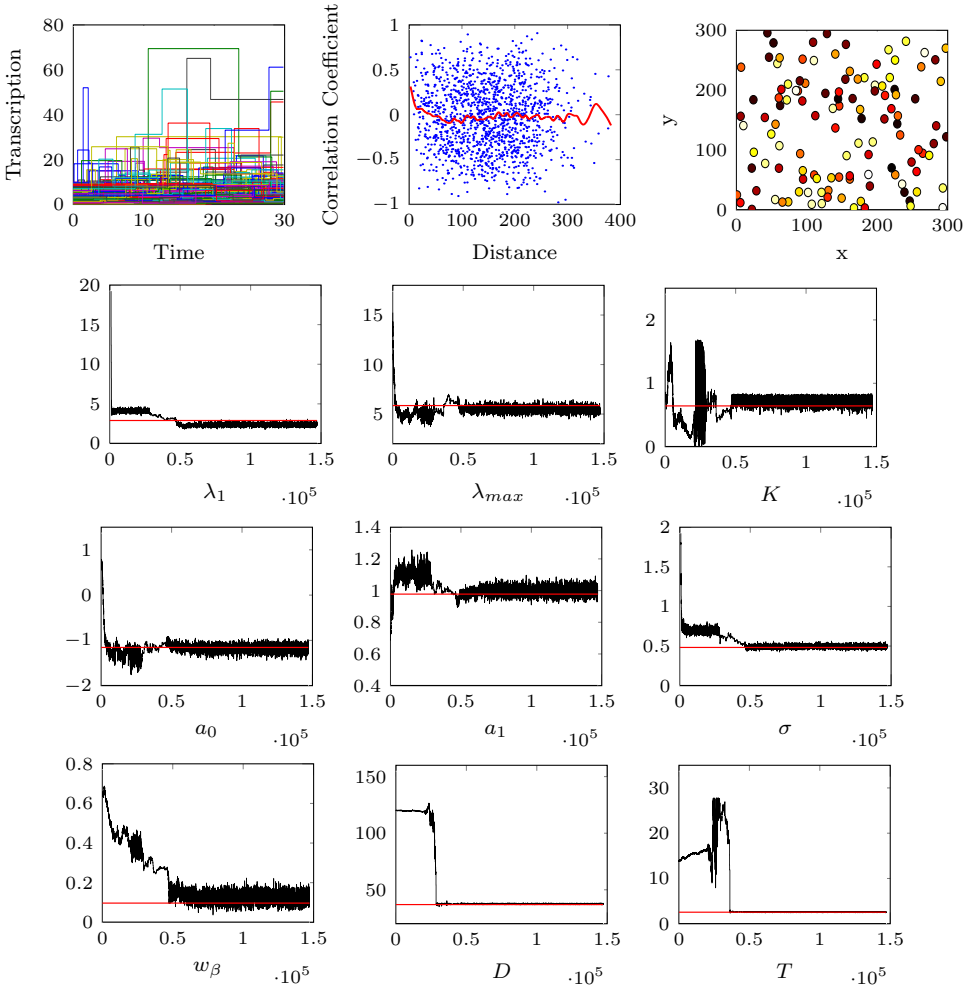


Figure 7.20: Illustrative example of the convergence for spatial transcriptional model 3. Top row shows the data with first plot, the transcriptional profiles, second plot the correlation distance relationship and spatial location in the third plot. The remaining plots, show the traces of the Markov chains for estimating the parameters of spatial model 3. True values are indicated by the red line.

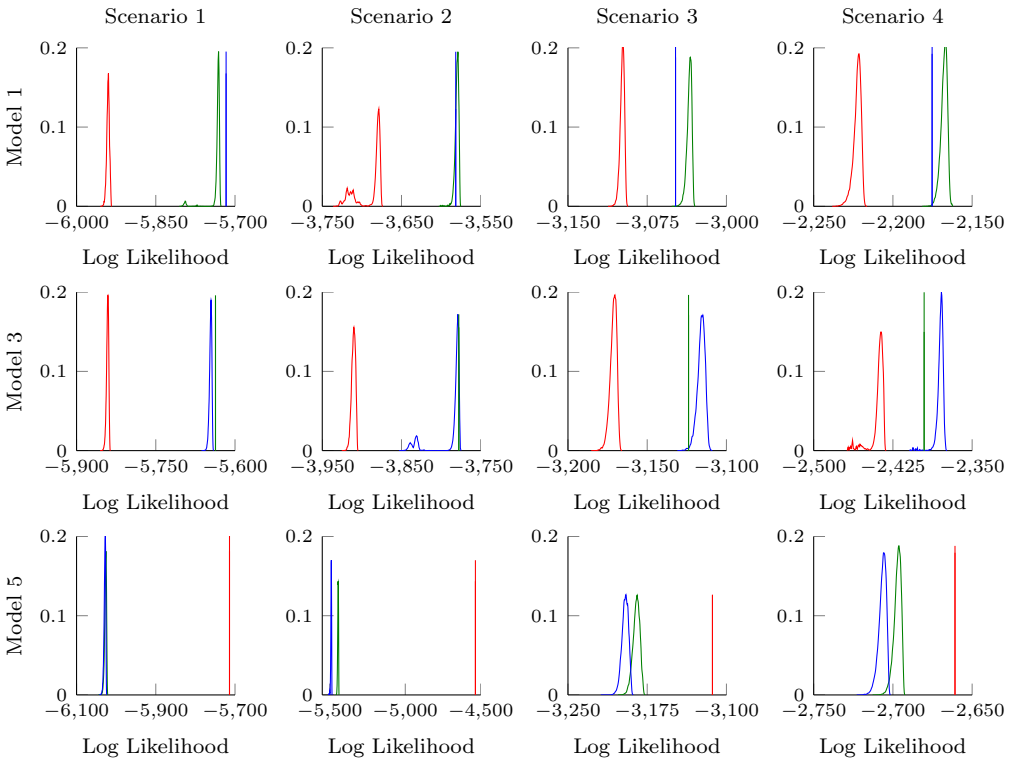


Figure 7.21: Likelihood comparison of different transcriptional models. Each column corresponds to a different data scenario described in the main text. The first row, is obtained from data simulated from model 1, with true likelihood value shown by the blue line, the corresponding densities of the "wrong" likelihood calculated from the MCMC runs by targeting the posterior of the wrong model are shown by the green (model 3) and red (model 5) curves. Similarly, the second row corresponds to data simulated from model 3, with true likelihood shown by the green line and "wrong" likelihoods shown by the blue (model 1) and red (model 5) curves. The third row shows data simulated from model 5 with true likelihood given by the red line and "wrong" likelihoods shown by the blue (model 1) and green (model 3) curves.

this, we ran a number of simulations where inference was performed on the wrong model.

In general, we found that even though inference was based on the wrong model, convergence was still seen in the Markov chains for the same data scenarios as those discussed in Section 7.6.1. Unsurprisingly, the posterior estimates of the parameters under the wrong model differed to the true parameter values under the true model, most notably, the strength parameter K and threshold parameters T and D had significant deviation. Thus, if one identified the wrong model, one would also misidentify the key spatial components T and D. In order to compare the different models, we investigated the behaviour of the likelihood calculated from the MCMC

runs and compared to the likelihood of the true model with true parameters. It was found that transcriptional model 5 was most distinguishable, with the true likelihood under model 5 being significantly larger than the likelihood calculated under either models 1 or 3 (Figure 7.21). In contrast, models 1 and 3 were difficult to distinguish with similar likelihood values (or in some cases, larger likelihood values under the wrong model), except in the case of very good data resolution (large number of switches per series, and a very connected tissue where each cell has a large number of connected neighbours), which is shown in Figure 7.21.

Consequently, except in the case of very good data resolution, one cannot distinguish between transcriptional models 1 and 3. However, if data behaved more similarly to model 5, this is likely to be detected. Given the unidentifiability between models 1 and 3, it means that one is not always able to distinguish between a linear spatial dependence or an exponential dependence. Biologically, this could relate to distinguishing between juxtacrine and paracrine signalling since the diffusive nature of paracrine signals precludes an exponential behaviour and the direct cell contact mechanism of juxtacrine signalling may suggest a linear dependence.

7.7 Discussion

This chapter has given a flavour of the various approaches one can take for spatio-temporal model building of single cell transcription. Unlike the methodology for temporal dynamics, spatio-temporal models will be highly specific to the gene of interest. To our knowledge, this is the first attempt at quantifying single cell transcription within a spatial domain in a parametric, semi-mechanistic approach. Fully mechanistic models exist for specific systems including the mammalian clock (Ananthasubramaniam et al., 2014), somite formation (Terry et al., 2011), gene/protein interactions within *Drosophila* (Li and Chen, 2009) and on a smaller scale, the opening and closing of gap junctions (Vogel and Weingart, 1998; Paulauskas et al., 2009). However, these models assume a much greater knowledge of the signalling mechanisms and are not easily implemented in an exploratory framework. In contrast, the work in this chapter, although specific to Prolactin expression within the mammalian pituitary, has provided an exploratory framework for analysing and extracting spatial information of back calculated transcriptional dynamics. Moreover, we have shown, how different parametric models can be constructed in order to incorporate a spatial dependence without assuming any knowledge of the true signalling process.

The exploratory analysis revealed evidence of short range synchronicity in the back calculated transcriptional dynamics of fully mature pituitary tissue. Interestingly, from these analyses, we see changing spatial dependence over pituitary development with no synchronisation evident in immature pituitary tissues. Thus, construction of spatio-transcriptional models were based on evidence observed in the adult tissues. These parametric models of the transcriptional process have allowed us to investigate different properties of the signalling mechanisms. Moreover, we have focussed attention on investigating the properties and limitations of these models to find the scenarios in which full model identifiability can be achieved. It becomes important to couple these spatio-transcriptional models with the organisational structure found in Chapter 6, since we found this was able to explain some properties of the observed spatial dependence.

Of course, many improvements to the constructed transcriptional models can be made, particularly if one wanted to incorporate a specific signalling mechanism. However, our simulations have shown that although parameter identifiability can be achieved with data of a similar resolution to the observed data, model identifiability is likely to be difficult. It therefore remains to apply these models to real data to investigate if signalling mechanisms can be extracted. We have identified the basic requirements in order to to do this, in the hope that future data will have sufficiently good resolution to extract meaningful information. For instance, identifying a threshold distance, D could be indicative of juxtacrine (if $D =$ cell size), paracrine (if D small but greater than the average cell size) or long range signalling through another mechanism (D large). Coupled with this, if model 5 was seen to fit the data best, the threshold time T, may aid in the identification of the biological processes involved in signalling. For example, if D is small but T large, it may imply the signal is significantly delayed, possibly through a long series of intermediary reactions before affecting neighbouring transcriptional dynamics. Thus, although we have not applied these models to data, these simulations have helped identify the potential information these different models can yield.

Part IV

Summary

CHAPTER 8

FUTURE WORK, EXTENSIONS AND CONCLUSIONS

[A] quotation is a handy thing to have about, saving one the trouble of thinking for oneself, always a laborious business.

A.A. Milne, If I May

8.1 Future Work and Extensions

The data motivating this research provides an infinite number of possibilities for analysis. We have chosen to focus our attention to three key areas and discuss here the various possibilities for future extension.

At the fundamental level, one can consider adapting the measurement equation to incorporate further details of the measurement process. For instance, there is typically a limited dynamic range that fluorescence detectors can measure and the underlying process may be more accurately modelled by incorporating a threshold of the form,

$$Y = \begin{cases} \gamma_1 + \epsilon_1, & \text{if} \quad \kappa P \leq T_1 \\ \gamma_2 + \kappa P + \epsilon_2, & \text{if} \quad T_1 < \kappa P \leq T_2 \\ \gamma_3 + \epsilon_3, & \text{if} \quad \kappa P > T_2, \end{cases}$$

where ϵ_1, ϵ_2 and ϵ_3 are independent random components, possibly Gaussian so that $\epsilon_i \sim N(0, \sigma_i^2)$ and T_1 and T_2 are the limits of the detection range of the microscopes. Clearly, the complexity of the system increases significantly, with six extra parameters introduced, although in practice, one may have prior information about the parameters T_1, T_2, γ_1 and γ_3. In addition, the dependence of the measurement equation on the underlying population levels means that even under the LNA, the data likelihood becomes intractable and one would have to numerically integrate over the latent states either through a Gibbs sampler or a pseudo-marginal approach.

With regards to the temporal methodology, one can foresee further developments of the BDA approach to stochastic reaction networks. For instance, there are many stochastic reaction networks that can be expressed as a sequence of conditionally independent linear subsystems and it would be interesting to develop the theoretical properties of the accuracy of the approximation approach, in particular, in comparison to other approximations found in the literature.

In order to further develop the analysis of the spatial organisation of lactotroph cells, further data is required especially if the aim is to provide more robust conclusions regarding the development of tissue architecture. In addition, if one had data on other tissue features such as capillary networks or the position of other cell types, one may be able to incorporate additional information into the modelling approach to gain more insightful analyses.

The final and perhaps most significant place for future work is the development of a spatio-temporal model for transcription. We have investigated how one may incorporate a spatial dependence in the transcriptional function that reproduces much of the observed spatial dependence in the mature pituitary tissues. We have also investigated the data resolution required to elicit this information from single cell time series. In order to develop these models further, one would aim to apply them to the observed data through the techniques discussed in Chapter 7. Furthermore, one may wish to develop the models in a more mechanistic way with the information elicited. For example, if one can distinguish between the different forms of spatial signalling, this may then be incorporated explicitly into the model. One example is the distinction between paracrine and juxtacrine signalling.

8.2 Conclusions

To conclude, as with all research, there are many things one can extend, adapt or reformulate. In particular, with the ever evolving experimental framework, statistical procedures will need to be constantly updated in order to extract the most information for any given dataset. This thesis has been motivated by data obtained from a green fluorescent reporter protein imaged within single cells and as such, all methods and analyses have been tuned towards this type of data.

We believe that this thesis has achieved three main aims. Firstly, we have developed and applied a robust statistical methodology for back-calculating to the transcriptional dynamics from single cell time series data observed through light microscopy. The novelty of the multi-state stochastic switch model enables one to back-calculate to the transcriptional level without assuming any prior knowledge of the activation mechanisms of the gene promoter and can thus be more widely applied than typical switch models given in the literature. Secondly, we have demonstrated how spatial point processes may be used to model the spatial organisation of cells within intact tissue over changing environmental conditions. Although standard point processes fit the data reasonably well, we have seen that incorporating further structure may elicit further insight into the architecture of the tissue. Finally, we have laid the groundwork for formulating the multi-state switch model in a spatial domain. We have seen how different semi-parametric signalling mechanisms can be incorporated to reproduce the observed spatial dependencies. Moreover, through simulations we have seen how the spatial organisation itself can explain some of the observed spatial transcriptional behaviour. Thus, this thesis provides the temporal and spatial analysis of single cell gene expression that has been incorporated into the formulation of a spatio-temporal framework for future analysis.

Part V

Appendices and Bibliography

APPENDIX A

SUPPLEMENTARY REVIEW MATERIAL

A.1 Exact Inference Approaches

In this section, we review the various approaches to performing inference on exact stochastic reaction networks mentioned in Section 2.2.

Boys et al. (2008) describe a block update method for proposing a latent path. Considering each interval of observed data separately, $[t_i, t_{i+1})$, the algorithm proceeds by firstly proposing the number of reactions n_j that have occurred in the interval for each reaction j. Since it is known that the number of reactions occurring in an interval is Poisson (Kurtz, 1971), a proposal is defined by the difference of two Poisson random variables. The next step in the update is to propose the timings of the events. This is achieved by sampling from the approximate inhomogeneous Poisson process with rate given by,

$$\lambda_j(t) = (t_{i+1} - t)h_j(\mathbf{x}(t_i), \theta_j) + (t - t_i)h_j(\mathbf{x}(t_{i+1}), \theta_j), \tag{A.1}$$

for each reaction j. Having obtained the number and timings of events, the proposed latent path $\mathbf{z}^*$ is defined and accepted with probability $\alpha = \min(1, A)$, where,

$$A = \frac{\frac{d\mathbb{P}(z^*|\mathbf{x})}{d\mathbb{Q}} \prod_{j=1}^{J} q(n_j^*|n_j)p(n_j^*)}{\frac{d\mathbb{P}(z|\mathbf{x})}{d\mathbb{Q}} \prod_{j=1}^{J} q(n_j|n_j^*)p(n_j)}, \quad n_j \sim Pois\left(\frac{h_j(\mathbf{x}(t_i), \theta_j) + h_j(\mathbf{x}(t_{i+1}), \theta_j)}{2}\right),$$

where $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$ is the Radon-Nikodym derivative associated with the approximating Poisson process to the true process. Boys et al. (2008), applied this method to the Lotka-

Volterra model, simulated with large molecular numbers and small rate parameters (i.e. systems with little intrinsic noise). Not only did they apply the method to discretely sampled data but also to scenarios with entire species unobserved. Parameter estimation was shown to perform well in the Lotka-Volterra model where the number of predators was assumed to be completely unobserved. However, although stated otherwise, the applicability of the method to low molecular counts is questionable. In particular, when $\mathbf{x}(t_i) = \mathbf{x}(t_{i+1}) = 0$ for any interval $[t_i, t_{i+1}]$, the ratio of prior probabilities becomes ill-defined.

In a similar way to the method described above, Amrein and Künsch (2012) propose new latent paths in two steps. First, the number of reactions is proposed with associated reaction types sampled with probability proportional to their hazard rate. Secondly, the reaction timings are defined by sampling the distance between successive reactions from a Dirichlet distribution. This algorithm is coupled with a filtering technique to generate reasonable starting values, which greatly improves the efficiency of the algorithm. In addition, the authors assume unknown measurement error, which turns out to be essential in the estimation process. However, the methods showed greatly reduced accuracy in data poor scenarios, where only a subset of species were observed, and are found to be computationally cumbersome.

The above methods of latent sampling have been incorporated into a natural Bayesian framework since parameter estimation can easily be computed given a full sample path due to the existence of conjugate priors. Daigle et al. (2012) have developed a frequentist approach to parameter estimation of biochemical systems, with their MCEM^2 (Monte Carlo Expectation Maximisation with Modified Cross-Entropy Method) algorithm. The EM (Expectation Maximisation) algorithm is used to iteratively update parameter estimates θ,

$$\begin{aligned}
\hat{\theta}^{(n+1)} &= \arg\max_{\theta} \left(\mathbb{E}(\log f_\theta(x, z) | \mathbf{x}, \hat{\theta}^n) \right) \\
&= \arg\max_{\theta} \left(\sum_{z \in Z(y)} f_z(z|\mathbf{x}, \theta) \log f_\theta(x, z) \right) \\
&\approx \arg\max_{\theta} \left(\sum_{k=1}^{K} \underbrace{\mathbb{I}[z_k^{(n)} \in Z(y)]}_{\substack{\text{simulated} \\ \text{trajectories}}} \log f_\theta(x, z_k^{(n)}) \right)
\end{aligned}$$

where the last expression is the Monte Carlo extension to the EM algorithm. Here the simulated trajectories are conditional on the observed data. Thus the issue is to sample trajectories with parameter vector θ that are consistent with the data. This

is achieved via rare event simulation techniques such as cross-entropy (Rubinstein, 1997). Daigle et al. (2012) compared the MCEM^2 algorithm to the block update method of Boys et al. (2008) and found that the MCEM^2 took significantly longer to run. However, they did note that although the block update loses accuracy as molecular numbers become small the MCEM^2 remained accurate.

An alternative approach to sampling the full trajectory is via a particle marginal Metropolis Hastings (PMMH) method. This has been suggested in Andrieu et al. (2009) and studied in more detail in Golightly and Wilkinson (2011). Similarly to the MCEM^2 of Daigle et al. (2012), the PMMH method is a way of simulating trajectories from the exact model via a stochastic simulation algorithm whilst ensuring these simulated latent paths are consistent with the observed data. This is achieved via sequential Monte Carlo (SMC) methods, which Golightly and Wilkinson (2011) find to be computationally burdensome. Moreover, the method performs badly in low/no measurement error scenarios.

Two recent examples that apply their methods to real data are the delayed acceptance MCMC method of Golightly et al. (2014) applied to epidemic data and the dynamic prior propagation method of Zechner et al. (2014) who model an artificially controlled gene expression system in yeast. The delayed acceptance method of Golightly et al. (2014) is again an application of particle MCMC methods. However, in order to decrease the computational cost, sample paths are first proposed under a fast approximation and only if these are accepted is a sequential Monte Carlo scheme used to estimate the true latent states. In this way, their algorithm avoids computing proposals under the true likelihood that are likely to be rejected. This method was applied to several synthetic reaction networks and also to a real epidemic dataset, however, when applied to data, the authors assumed there to be no measurement error.

The dynamic prior propagation method of Zechner et al. (2014) is based on a hierarchical Markov model and is applied to multiple trajectories of single cell gene expression time series data. The authors construct a hierarchical model over the different cells such that some parameters are assumed constant across the dataset, whilst others are cell specific. In this way, the authors construct a marginal reaction network over the rate constants that vary between cells, and instead only infer sufficient statistics of the distribution of these rate constants rather than each individual cell specific constant. The authors successfully applied their approach to single cell time course measurements under Gaussian or log-Gaussian white noise measurement error. Although described as a scalable approach, no indication of

computational cost is given. Moreover, it is clear the method scales well as the number of individual cells increases but it is less clear how well the methods will scale as the number of observations per cell increases, particularly since it relies on an SMC update through stochastic simulation of the latent states.

One further approach to inference on the exact MJP has been achieved through approximate inference techniques. In these scenarios one continues to work with the exact stochastic system but obtains only an approximation to the true posterior density. Examples include ABC (Approximate Bayesian Computation, Beaumont et al. (2002)) and Variational Bayes. For instance, the basic ABC algorithm takes the following form,

1. Generate a candidate parameter set θ^* from a prior distribution.
2. Simulate a dataset $\mathbf{X}_{\theta^*}$ using the parameter set θ^*.
3. Evaluate the distance, $d(S(\mathbf{X}_{\theta^*}), S(\mathbf{Y}))$, where $\mathbf{Y}$ is the observed data, $S(.)$ is a summary statistic of the datasets and d is some metric.
4. If $d(S(\mathbf{X}_{\theta^*}), S(\mathbf{Y})) < \epsilon$, for some predefined threshold ϵ, accept θ^*, otherwise reject.
5. Repeat.

In this way, one obtains samples θ^* of an approximate posterior density. If sufficient statistics are available, as in the stochastic autologistic model of (Drovandi and Pettitt, 2011), and are used as the summary statistics in step 3, as $\epsilon \to \infty$, the true posterior is recovered. If however, sufficient statistics are not available, one has to choose the summary statistics. Fearnhead and Prangle (2012) develop a semi-automatic approach to choosing the summary statistics when sufficient statistics are unavailable and apply their methods to simulated data from the stochastic Lotka-Volterra model. Alternative implementations include the gene network models of Lillacci and Khammash (2013) who choose d to be the Kolmogorov distance between the empirical CDFs of the datasets $\mathbf{X}_{\theta^*}$ and $\mathbf{Y}$, and can be viewed as a distribution matching approach.

An alternative approximate inference technique can be achieved through Variational Bayes, where one approximates the posterior $f(\theta|\mathbf{y})$, by some density q, where q is chosen optimally from a certain family of distributions. For example, Opper and Sanguinetti (2010), infer the logics of two competing transcription factor interactions within a reduced MJP and find q to optimise the Kullback-Leibler divergence where q is restricted to the family of Markov jump processes.

A.2 Alternative Approximations

As discussed in Section 2.3.4, there are many different ways in which one can approximate a stochastic reaction network beyond the LNA, CLE or BDA. Here, we present a review of the different approximations that can be found in the literature.

In general, approximations to stochastic reaction networks can be categorised into two main approaches (Kazeev et al., 2014). Firstly, there are the approximations obtained asymptotically, which include the macroscopic rate equation, the CLE and the LNA as already discussed in detail in Section 2.3.2. Other examples include the Moment Closure approximations (MA), quasi-steady state approaches (QSSA) and extensions to the LNA. As the name suggests, Moment Closure approximations consider only the moments of the solution to the master equation rather than the full transition density. In the case of unimolecular reactions, expressions for the moments can be solved exactly. However, for systems with at least one bimolecular reaction, the equations for the moments form an infinite hierarchy where each equation depends upon higher order moments. Thus moment closure approximations consist of truncating this infinite hierarchy to obtain a closed system of equations. Moment approximation methods (Gomez-Uribe and Verghese, 2007; Ferm et al., 2008; Ullah and Wolkenhauer, 2009) have been used for stochastic simulations and to study properties of the noise of a given system. However, the MA approach is only valid under certain conditions (Schnoerr et al., 2014) and the truncation level can be highly influential.

The quasi-steady state assumption (Rao and Arkin, 2003) enables a simplification of stochastic reaction networks by eliminating reactions of fast dynamics that are computational expensive to simulate. This is achieved by partitioning the system into fast and slow reactions where the transition density of the fast reactions conditional on the slow reactions is assumed to be constant in time. Thus, the system reduces to consist of only slow reactions whose hazard rates will no longer be governed by the law of mass action. An explicit example is given by the Michaelis-Menten dynamics of substrate-enzyme interactions, described by the three reaction network,

$$E + S \underset{k_0}{\overset{k_1}{\rightleftharpoons}} ES, \quad ES \xrightarrow{k_2} P,$$

and can be reduced to the single non-linear reaction,

$$S \to P,$$

with corresponding reaction rate, $h(s, \theta) = k_2 es / \left(\frac{k_0 + k_2}{k_1} + s \right)$, where e is the total number of enzymes and s is the number of substrates. Other examples include models for gene expression, where one approximates the activity of the promoter through a quasi-steady state approximation. Note that this is different to the gene expression model considered in this thesis. Although promoter activity is modelled only through the transcriptional function, $\beta(t)$, our model remains linear and does not assume any particular form of the promoter activity and is in fact one of the quantities we may wish to infer. This is important since the quasi-steady state approach has been shown to incorrectly model the intrinsic noise of the system and moreover can be shown to produce qualitatively misleading behaviour (Thomas et al., 2012).

The linear noise approximation was introduced in Section 2.3.2 and can be viewed as a system size expansion of the master equation. The LNA has many advantages for an approximation to be used for inference, however it is shown in Section 2.4, that it has reduced simulation accuracy compared to the CLE and BDA. Consequently there are several extensions to the LNA that have been discussed in the literature for increased accuracy of stochastic simulations. First, there are the effective mesoscopic rate equations (EMREs) of Grima (2010), which are derived through a system size expansion of the master equation. This expansion explicitly includes terms of one order higher than the expansion leading to the conventional LNA. As a consequence, the EMRE then includes a "correction" term that accounts for the discrepancy between the mean of the system and the macroscopic ODE. Thus for linear systems (systems consisting of at most first order reactions), the LNA and EMRE will coincide.

Secondly, there is the slow-scale LNA (ssLNA) of Thomas et al. (2012). As with the quasi-steady state approach, this method is motivated by the idea of timescale separation, where reactions can be partitioned into fast and slow reactions. The ssLNA is derived by marginalising the multivariate Gaussian transition density of the LNA to obtain the reduced LNA for only the slow reactions. The authors compare this approach to the LNA derived on a reduced QSSA stochastic system to find significant differences in the behaviour. However, one should note that typically, under the ssLNA, one retains more parameters than under the QSSA and so if the aim is model reduction, the ssLNA may not be restrictive enough.

One final extension to the LNA is the recently derived conditional LNA (Thomas et al., 2014). Derived explicitly for gene regulatory networks, the main idea is to model the dynamics of promoter regulation exactly and then conditional on each pro-

moter state, approximate the population dynamics of the remaining species through the LNA. This results in an overall mixture distribution consisting of the sum of each of the conditional LNAs. This is an interesting approach that is able to capture bimodal behaviour that has been observed experimentally but cannot be captured through the conventional LNA. However, this approach relies on the assumption that promoter dynamics are regulated via unspecific effects, i.e. not by the gene products one models through the LNA. If one knew the regulatory mechanisms of the gene promoter, this approach could provide interesting analysis, particularly if cells were unsynchronised. However, for our application to the Prolactin gene, little is known about regulation and consequently, this approach is less amenable.

The second broad approach to approximating the exact MJP is through a truncation of the state space of the system. For example, the finite state projection (FSP) method (Munsky and Khammash, 2006, 2008) allows one to approximate the solution to the master equation to any prespecified degree of accuracy. Explicitly, letting A be the infinitesimal generator matrix such that the exact process satisfies the intractable system,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}(\mathbf{x},t) = A\,\mathbb{P}(\mathbf{x},t),$$

one can find a finite projection, $\mathbb{P}^{\mathrm{FSP}}$ satisfying,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}^{\mathrm{FSP}}(\mathbf{x},t) = A\,\mathbb{P}^{\mathrm{FSP}}(\mathbf{x},t),$$

such that,

$$\left|\begin{pmatrix}\mathbb{P}_J\\ \mathbb{P}_{J'}\end{pmatrix} - \begin{pmatrix}\mathbb{P}^{\mathrm{FSP}}\\ 0\end{pmatrix}\right| \leq \epsilon, \quad \text{and} \quad \mathbb{P}^{\mathrm{FSP}}(0) = \mathbb{P}_J(0),$$

where J denotes the set of states included in the projection $\mathbb{P}_J$. This approach has the advantage that one knows directly the prespecified error, which is introduced by the approximation. The FSP method has recently been adapted by quantised tensor trains (Kazeev et al., 2014) that reduce the computational cost and enables greater feasibility of these truncation methods. However, one still has to choose a truncation point and the methods will be expensive if the state space is large. Moreover, the system size is often unknown *a priori* and it is therefore unclear if the approach is feasible for the problem at hand.

Choi and Rempala (2012) also truncate the state space to implement a uniformisa-

tion method (Hobolth and Stone, 2009) to calculate the latent path. The method can be broken down into three parts. Consider time interval [0,t] with observed endpoints. Let P be the matrix of transition probabilities between states of the process X and define $R = \frac{1}{\mu}P + I$ to be the transition matrix for a new process. It is assumed that the state space can be truncated such that the dimension of P is bounded and that μ is the rate of the Poisson counting process N (the number of events/reactions occurring in $[0, t)$) such that the following identity is satisfied,

$$\mathbb{P}(X(t) = j | X(0) = i, N(t) = n) = R^n_{ij}.$$

Given the above identify it can be shown that,

$$\mathbb{P}(X(t) = j | X(0) = i) = \exp(tP)_{ij}.$$

The number of events occurring in an interval $[0, t)$ can therefore be sampled from the following,

$$\mathbb{P}(N(t) = n | X(t) = j, X(0) = i) = \frac{(\mu t)^n e^{-\mu t}}{n!} \frac{R^n_{ij}}{[\exp(tP)]_{ij}}.$$

Having done this, the type of events that have occurred is calculated iteratively,

$$z_1 \sim \mathbb{P}(z_1 = l | z_0 = i, z_n = j) = \frac{R_{il} R^{n-1}_{lj}}{R^n_{ij}},$$
$$z_2 \sim \mathbb{P}(z_2 = m | z_1 = l, z_n = j) = \frac{R_{lm} R^{n-2}_{mj}}{R^{n-1}_{lj}}, \ldots$$

and the timings of these events are proposed uniformly on $(0, t)$. In contrast to the FSP method, the transition density over discrete time intervals remains unavailable but one obtains a continuous time trajectory of the unobserved states. The main cost of the above method is in the eigen decomposition of the generator matrix, which could be of very high dimension depending on where one chose to truncate the state space.

APPENDIX B

TECHNICAL APPENDICES

B.1 Transition Densities for SRNs

We show here, through the example of independent birth and simple death how one can evaluate the transition density satisfying the corresponding master equation given below,

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}t} = \beta\mathbb{P}(x-1,t) \; + \; \delta(x+1)\mathbb{P}(x+1,t) - \; (\beta+\delta x)\mathbb{P}(x,t). \tag{B.1}$$

In the same way as Gardiner (1985) we solve the master equation by first introducing the generating function, $G(x,t) := \sum_{x=-\infty}^{\infty} z^x \mathbb{P}(x,t)$. We then have the following identities,

$$\frac{\partial G}{\partial z} = \sum_{x=-\infty}^{\infty} x z^{x-1}\mathbb{P}(x,t), \quad \frac{\partial G}{\partial t} = \sum_{x=-\infty}^{\infty} z^x \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}t}.$$

Thus, multiplying (B.1) by z^x and summing over x, we get,

$$\sum_{x=-\infty}^{\infty} z^x \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}t} = \beta \sum_{x=-\infty}^{\infty} z^x \mathbb{P}(x-1,t) + \delta_X \sum_{x=-\infty}^{\infty} (x+1) z^x \mathbb{P}(x+1,t) - \ldots$$
$$\cdots - \beta \sum_{x=-\infty}^{\infty} z^x \mathbb{P}(x,t) - \delta_X \sum_{x=-\infty}^{\infty} x z^x \mathbb{P}(x,t).$$

Rewriting the above in terms of the generating function, G, we obtain the following

PDE,

$$\frac{\partial G}{\partial t} - \delta_m(1-z)\frac{\partial G}{\partial z} = \beta(z-1)G. \tag{B.2}$$

Method of characteristics

Thus we have a first order PDE that can be solved via the method of characteristics. We introduce a new variable s and let $t = t(s)$, $z = z(s)$ and $G(z,t) = G(s)$. Thus,

$$\frac{\mathrm{d}G}{\mathrm{d}s} = \frac{dz}{\mathrm{d}s}\frac{\partial G}{\partial z} + \frac{\mathrm{d}t}{\mathrm{d}s}\frac{\partial G}{\partial t}$$

Solving the above on the characteristics $\frac{\mathrm{d}t}{\mathrm{d}s} = 1$ and $\frac{\mathrm{d}z}{\mathrm{d}s} = -\delta_m(1-z)$, we have,

$$\frac{dG}{ds} = \beta(z-1)G. \tag{B.3}$$

Firstly, solving the characteristics gives $t(s) = s$, and the second characteristic can be solved as follows,

$$\begin{aligned}
\frac{dx}{ds} &= -\delta(1-z) \\
z(s) &= 1 - Ae^{\delta s} \\
z(s) &= 1 - (1-z_0)e^{\delta s}
\end{aligned}$$

where $z(0) = z_0$ and thus,

$$z_0 = 1 - (1-z)e^{-\delta_X t}. \tag{B.4}$$

We can now solve the ODE (B.3) along the characteristics,

$$\begin{aligned}
G(s) &= G_0 \exp\left(\int_s \beta z - \beta \ \mathrm{d}s \right) \\
&= G_0 \exp\left(\int_s \beta(-1+z_0)e^{\delta s} \ \mathrm{d}s \right) \\
&= G_0 \exp\left(\tfrac{\beta}{\delta}(z_0-1)e^{\delta s} + B \right)
\end{aligned}$$

where, $G(0) = G_0$, which gives, $B = -\frac{\beta}{\delta}(z_0 - 1)$. Consequently,

$$G(z,t) = G_0 \exp\left(\tfrac{\beta}{\delta}(1-z)e^{-\delta t} + \tfrac{\beta}{\delta}(z-1) \right) \tag{B.5}$$

subject to initial condition $G_0 = G(z_0, 0) = z_0^{x_0}$. Using (B.4) we have,

$$z_0^{x_0} = \Big(1 - (1-z)e^{-\delta t}\Big)^{x_0}.$$

Consequently,

$$\begin{aligned} G(z,t) &= \Big(1 - (1-z)e^{-\delta t}\Big)^{x_0} \exp\Big(\tfrac{\beta}{\delta}(1-z)e^{-\delta t} + \tfrac{\beta}{\delta}(z-1)\Big) \\ &= \Big(1 - (1-z)e^{-\delta t}\Big)^{x_0} \exp\Big(-\tfrac{\beta}{\delta}(1-e^{-\delta t})\Big) \exp\Big(\tfrac{\beta}{\delta} z(1-e^{-\delta t})\Big) \\ &= \exp\Big(-\tfrac{\beta}{\delta}(1-e^{-\delta t})\Big) \Big(1 - (1-z)e^{-\delta t}\Big)^{x_0} \exp\Big(\tfrac{\beta}{\delta} z(1-e^{-\delta t})\Big) \end{aligned} \tag{B.6}$$

Expanding the above in terms of a power series we get,

$$\begin{aligned} G(z,t) = \exp\Big(-\tfrac{\beta}{\delta}(1-e^{-\delta t})\Big) \sum_{x=0}^{\infty} \sum_{k=0}^{x_0} \binom{x_0}{k} (e^{-\delta t})^k (1-e^{-\delta t})^{x_0-k} \times \dots \\ \times \Big(\tfrac{\beta}{\delta}(1-e^{-\delta t})^{x-k}/(x-k)!\Big) z^x \end{aligned} \tag{B.7}$$

Equating terms, we can then obtain the transition density;

$$\begin{aligned} \mathbb{P}(x,t) &= \sum_{k=0}^{\min(x,x_0)} \binom{x_0}{k} \pi^k (1-\pi)^{x_0-k} \frac{\lambda^{x-k} e^{-\lambda}}{(x-k)!} \\ \lambda &= \frac{\beta}{\delta}(1-e^{-\delta t}), \\ \pi &= e^{-\delta t} \end{aligned}$$

which is a convolution of a Poisson and a binomial pdf. The above result was derived assuming time independent kinetic rates, this is not a necessary condition, and in particular in order to extend the above result, λ and π will instead satisfy the following ODEs,

$$\begin{aligned} \frac{\mathrm{d}\lambda}{\mathrm{d}t} &= \beta(t) - \delta(t)\lambda, & \lambda(0) &= 0, \\ \frac{\mathrm{d}\pi}{\mathrm{d}t} &= -\delta(t)\pi, & \pi(0) &= 1. \end{aligned}$$

In the above example, the exact transition density can be obtained by solving the

PDE of the generating function. In a similar way, one can often obtain the moments of other stochastic reaction networks, even when the full transition density is intractable. However, there are many examples, where even the moments of the system are unavailable, including autoregulatory systems with reactions of the form,

$$X \to X.$$

B.2 Reparameterisation of the LNA

When performing inference on the gene transcription model under the LNA, it was found that reparameterising yielded significant improvement in the mixing properties of the Markov chains. To be explicit, the reformulation of the state space model under the linear noise approximation is given below. Throughout we let $\tilde{P} := \kappa P$ and $\tilde{\alpha} := \kappa\alpha$. Thus the observation equation is given by,

$$Y(t) = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} M(t) \\ \tilde{P}(t) \end{pmatrix} + \epsilon(t), \tag{B.8}$$

$$\epsilon(t) \sim N(0, \sigma_\epsilon^2). \tag{B.9}$$

Letting $X(t) = (M(t), \tilde{P}(t))^T$, recall that the LNA state equation takes the form,

$$\begin{aligned} X(t+\tau) &= e^{J\tau}X(t) + \Omega(\phi(t+\tau) - e^{J\tau}\phi(t)) + \eta(t+\tau), \\ \eta(t+\tau) &\sim N(0, \Sigma(t+\tau)) \\ \Sigma(t+\tau) &= \int_t^{t+\tau} [e^{-J(t+\tau-s)}B(s)][e^{-J(t+\tau-s)}B(s)]^T \, \mathrm{d}s, \end{aligned}$$

where now,

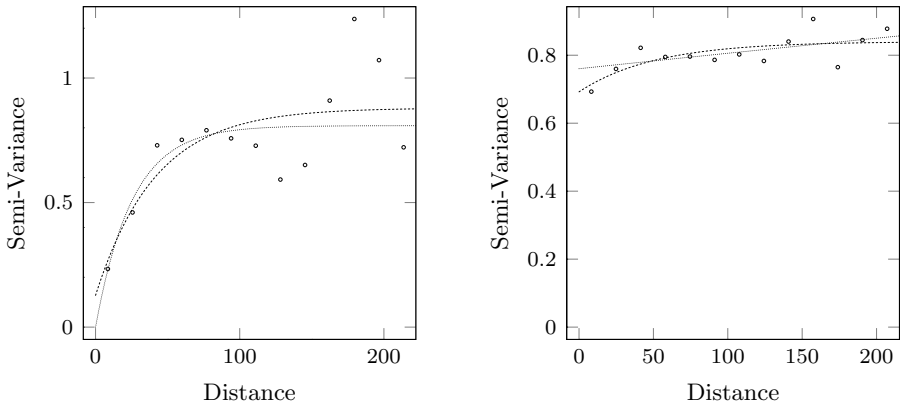
$$\begin{aligned} J &= \begin{pmatrix} -\delta_m & 0 \\ \tilde{\alpha} & \delta_p \end{pmatrix}, \\ B(s) &= \begin{pmatrix} \sqrt{\beta(t) + \delta_m M(t)} & 0 \\ 0 & \sqrt{\kappa}\sqrt{\tilde{\alpha}M(t) + \delta_p \tilde{P}(t)} \end{pmatrix}, \end{aligned}$$

and $\phi := (\phi_m, \phi_p)^T$, where ϕ_M and ϕ_P solve the following ODE system,

$$\frac{\mathrm{d}\phi_m}{\mathrm{d}t} = \beta(t) - \delta_m \phi_m(t),$$

$$\frac{\mathrm{d}\phi_p}{\mathrm{d}t} = \tilde{\alpha}\phi_m(t) - \delta_p\phi_p(t).$$

We note that the rescaling in the system in this way introduces a factor of κ in the variance-covariance matrix and is therefore the term in which identifies the scaling parameter when performing inference.



(a) (b)

Figure B.1: Example variograms with a) calculated from a spatial field with correlation and b) calculated from a spatial field with no spatial correlation.

B.3 Variogram

Let Z be a spatial random field defined continuously over some spatial domain D. The variogram is defined as the variance of the difference in Z evaluated at any two spatial locations, i.e.

$$2\gamma(x, y) = \mathrm{Var}(Z(x) - Z(y)), \quad \text{for } x, y \in D,$$

where the variogram is given by 2γ and γ is called the semi-variogram. Thus, a typical variogram (or more precisely semi-variogram) for a spatial random field subject to spatial correlation is shown in Figure B.1a). The range at which the semi-variogram asymptotes is the range of interaction and moreover, the asymptote itself indicates no spatial correlation at this distance. Consequently, a typical variogram indicating no spatial correlation is shown in Figure B.1b). All implementations of the variogram analysis have been performed with the **geoR** package (Ribeiro Jr and Diggle, 2001) in **R**.

APPENDIX C

FIGURES

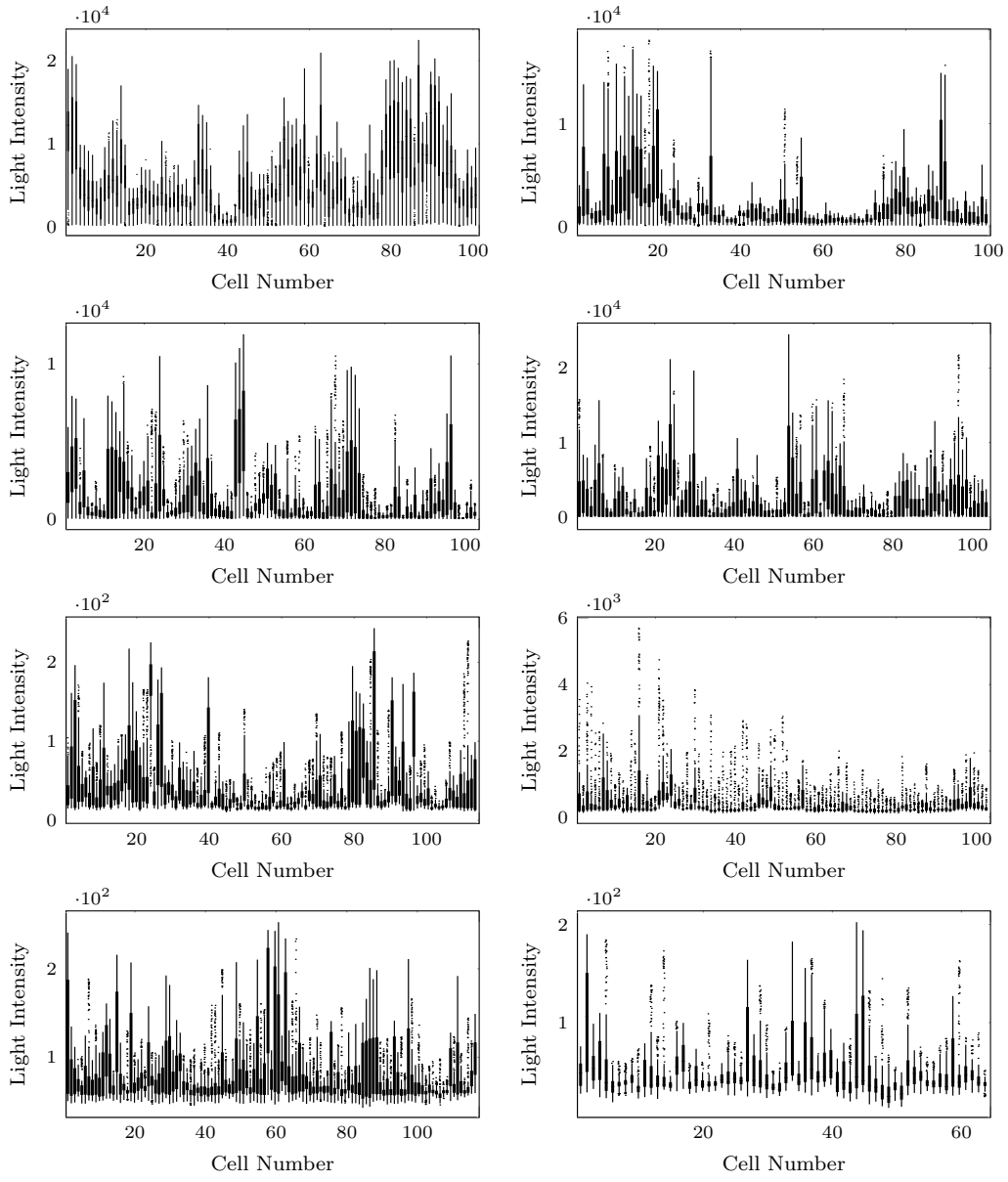


Figure C.1: Boxplots for each individual time series for all datasets consisting of A1-A4, P1-P2 and E1-E2.

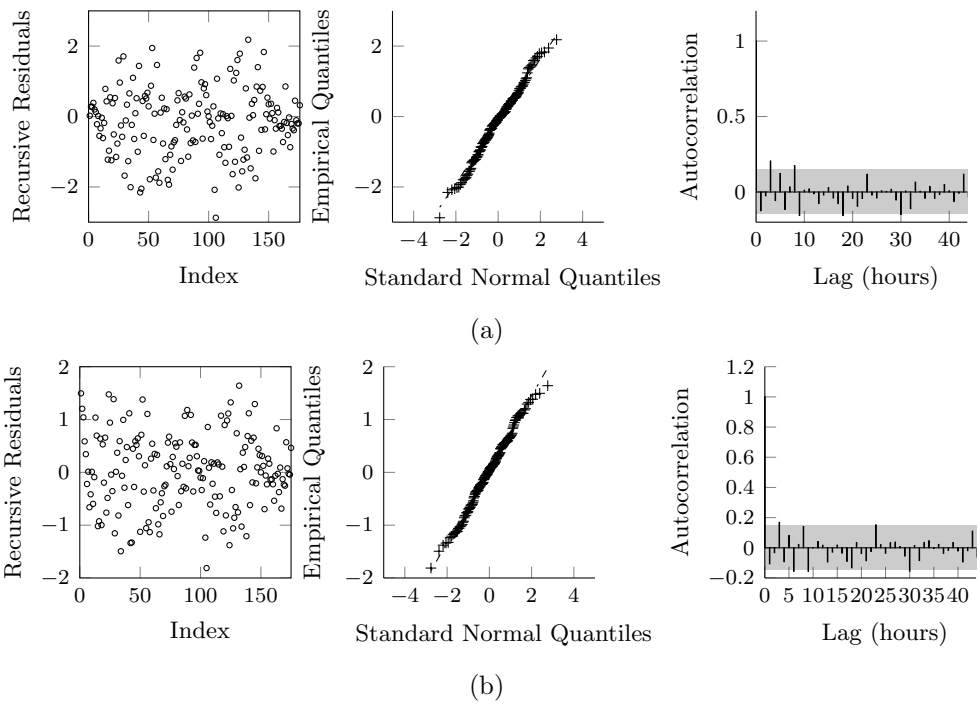
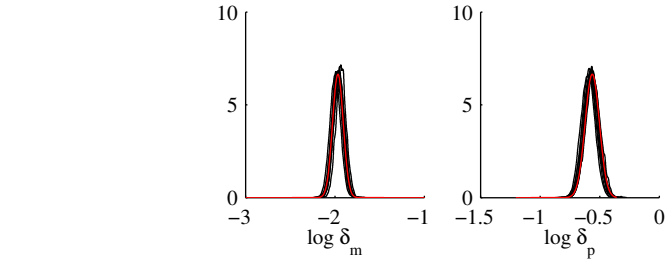
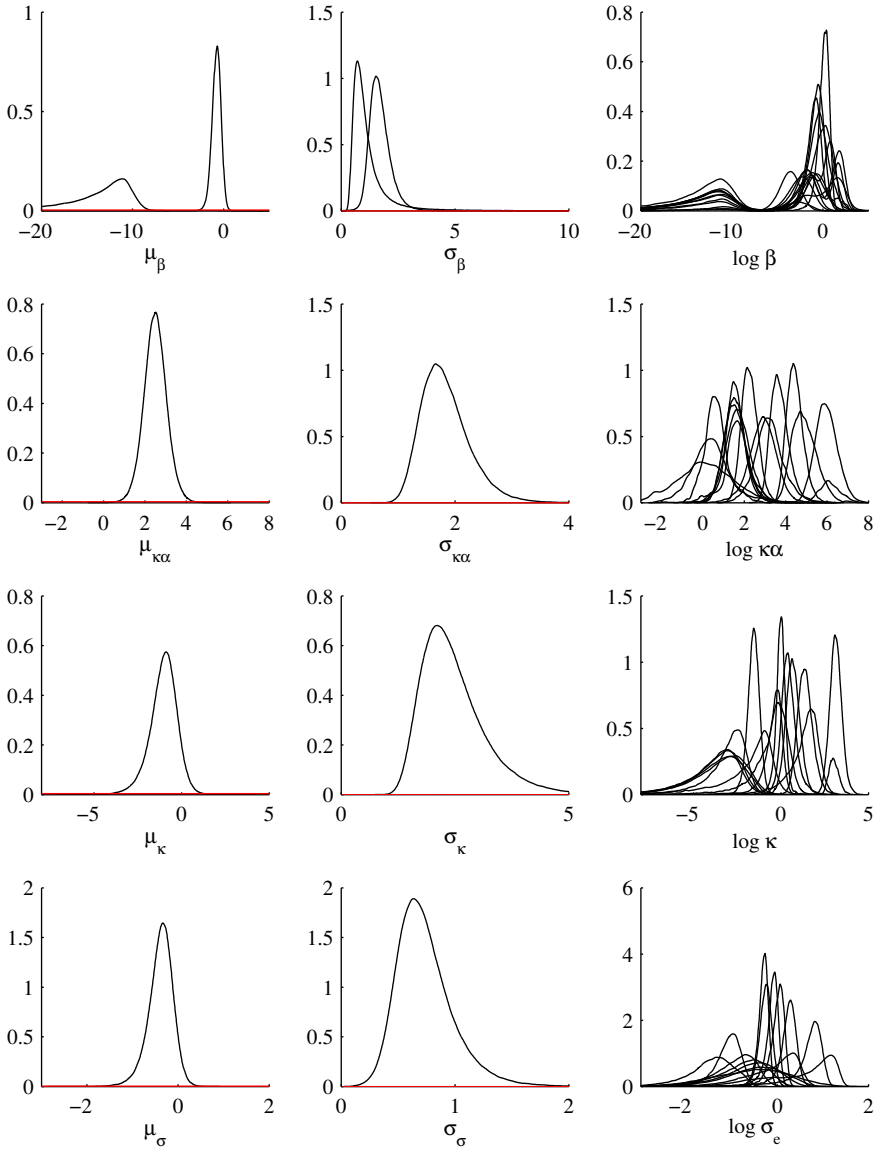


Figure C.2: Recursive residuals calculated for the time series shown in Figure 4.7 calculated at the posterior median estimates obtained under a) the LNA and b) the BDA. The left column shows the recursive residuals against index. The centre column gives a qq-plot of the residuals with no significant deviation from normality in both cases ($p < 0.05$ according to a Kolmogorov-Smirnov test for normality). The right column gives the autocorrelation of the residuals with the shaded region depicting the 95% envelopes of a white noise process.

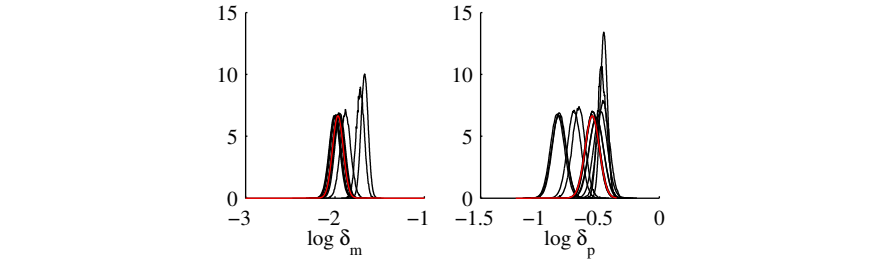


(a)

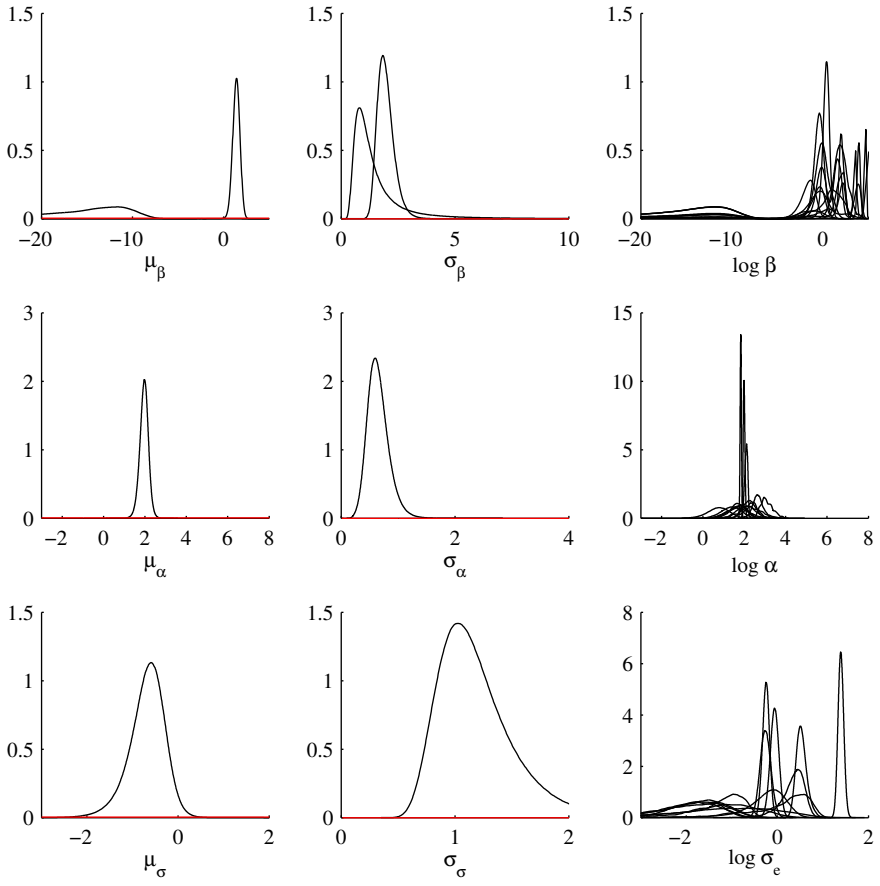


(b)

Figure C.3: Posterior densities estimated via the LNA on the subset of dataset P1. Red lines indicate the prior density. a) shows the posterior for the non-hierarchical parameters with b) showing all hierarchical parameters.



(a)



(b)

Figure C.4: Posterior densities estimated via the BDA on the subset of dataset P1. Red lines indicate the prior density. a) shows the posterior for the non-hierarchical parameters with b) showing all hierarchical parameters.

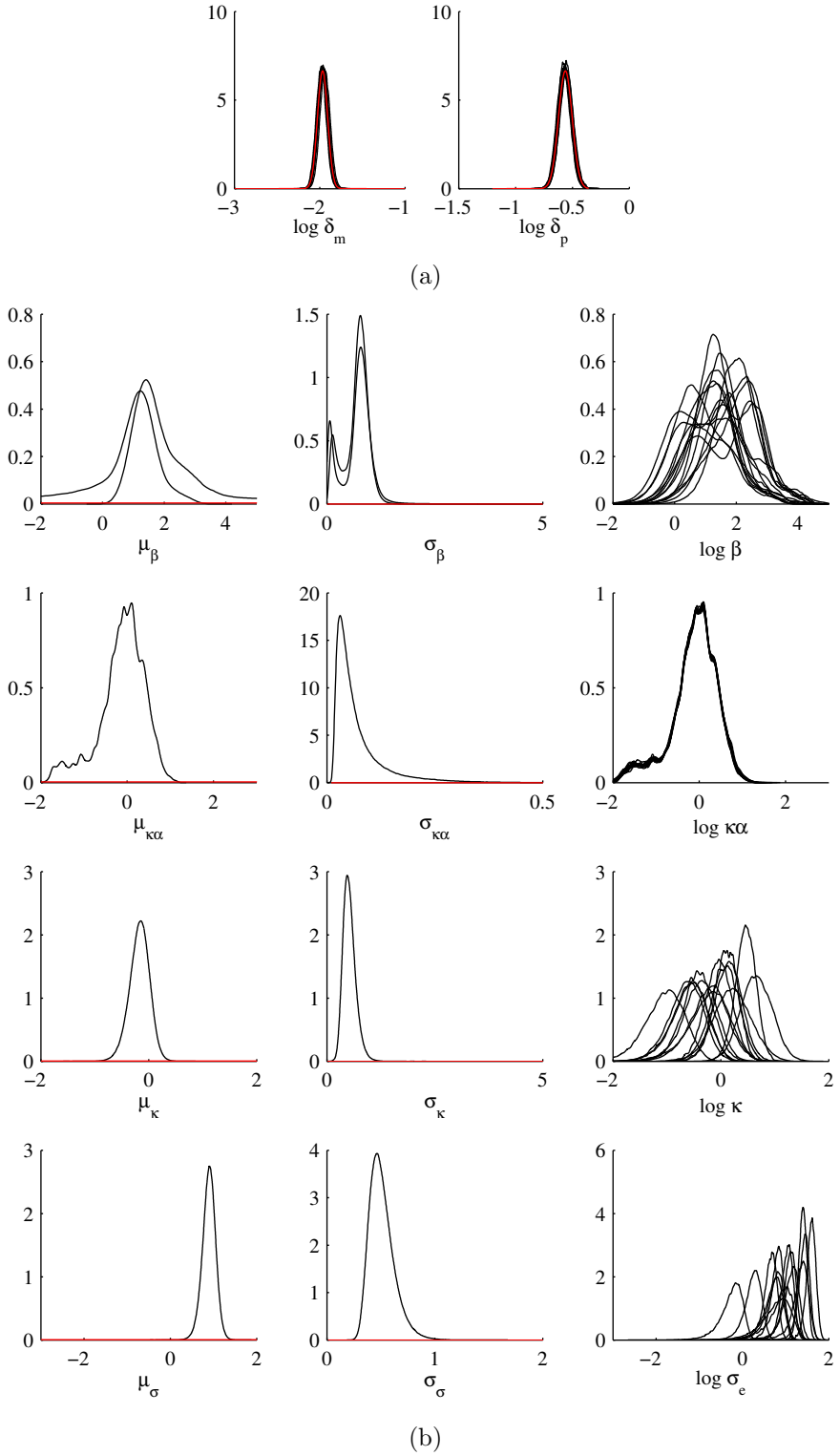


Figure C.5: Posterior densities estimated via the LNA on the subset of dataset A1. Red lines indicate the prior density. a) shows the posterior for the non-hierarchical parameters with b) showing all hierarchical parameters.

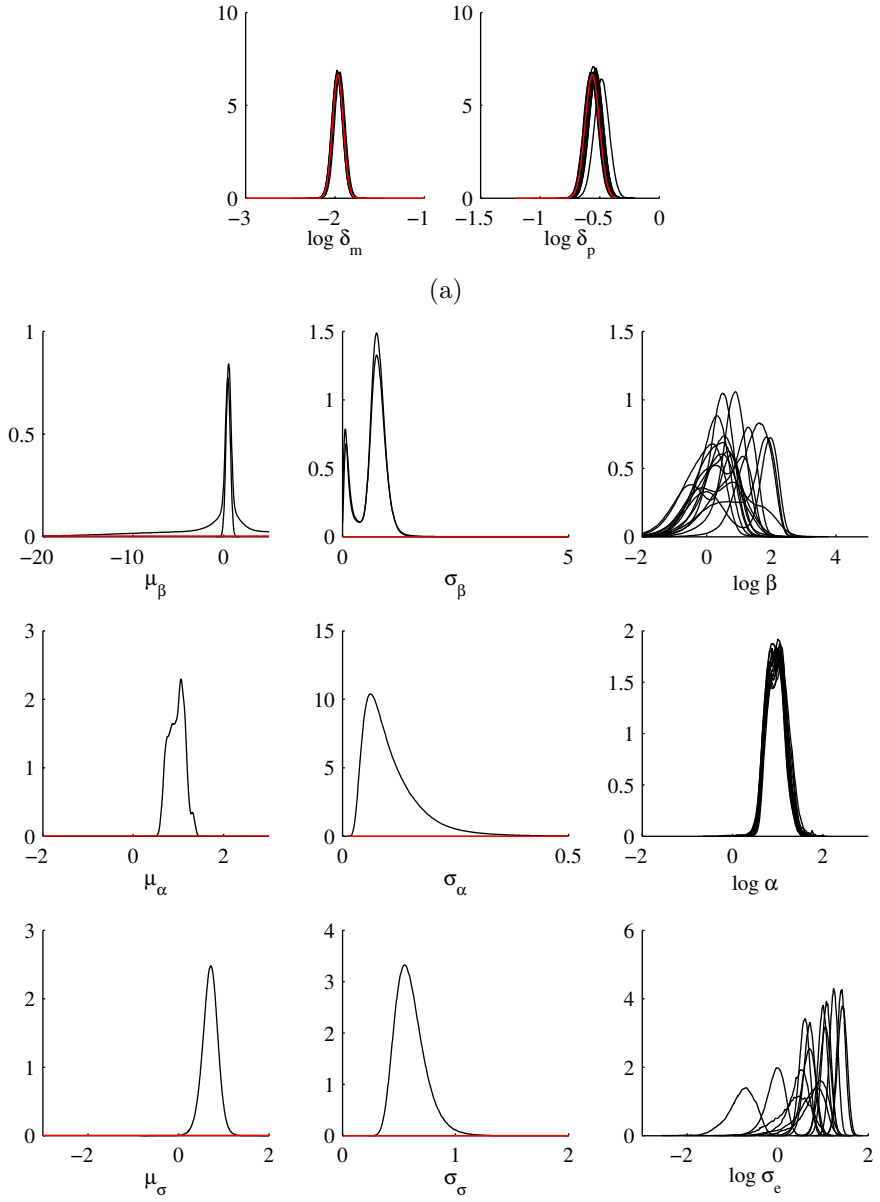


Figure C.6: Posterior densities estimated via the BDA on the subset of dataset A1. Red lines indicate the prior density. a) shows the posterior for the non-hierarchical parameters with b) showing all hierarchical parameters.

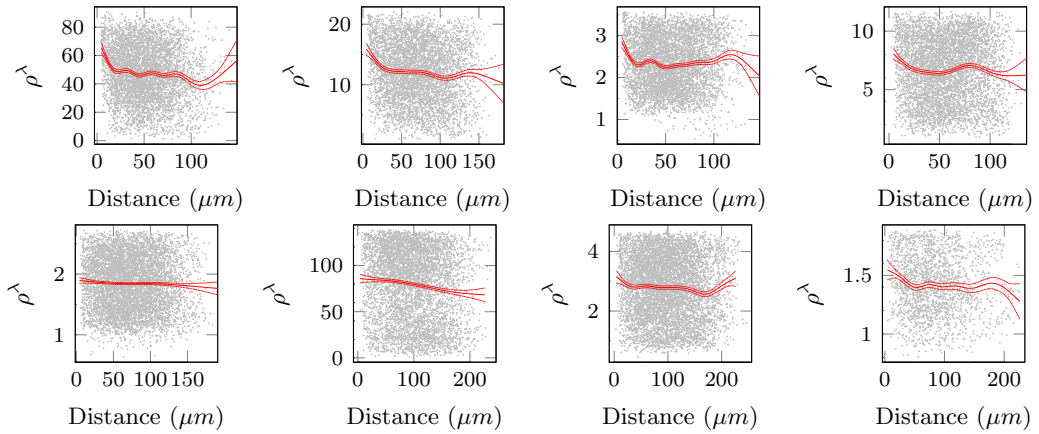


Figure C.7: The relationship between the pairwise correlation of the raw light intensity time series and their Euclidean distance. The correlation coefficient has been transformed via a Box-Cox transformation to fit a generalised additive model shown in red. The parameter λ has been chosen optimally for each dataset and thus, the y-scales are not comparable between datasets. To support Figure 7.1.

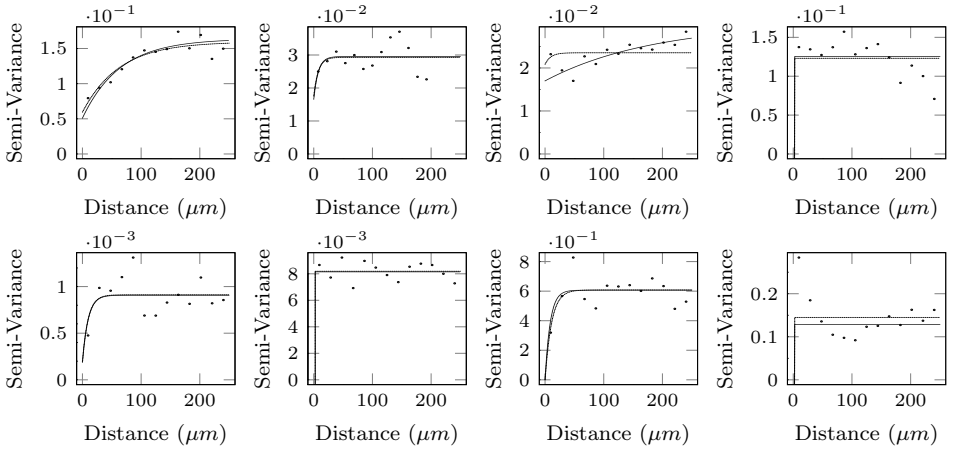


Figure C.8: Variograms for spatial Feature 1 with top row corresponding to datasets A1-A4, and bottom row corresponding to datasets P1-P2 and E1-E2. The dashed line corresponds to a ordinary least squares fit and dotted line corresponds to a weighted least squares fit.

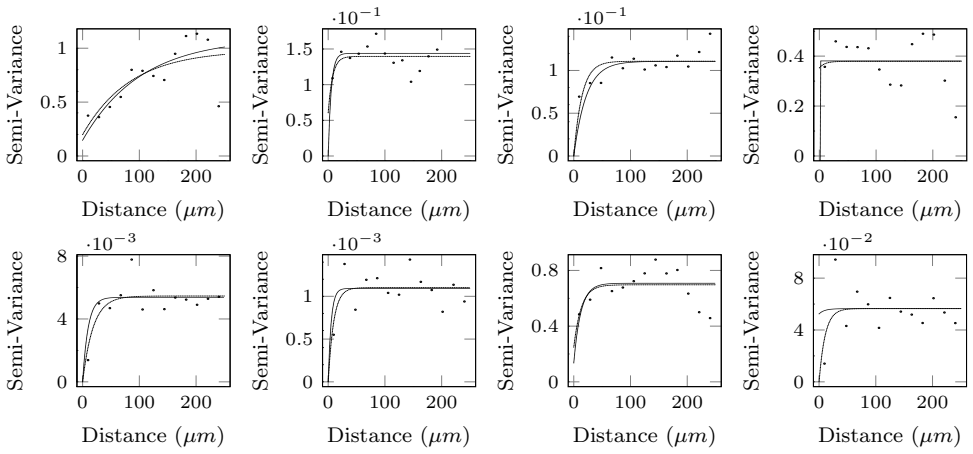


Figure C.9: Variograms for spatial Feature 2 with top row corresponding to datasets A1-A4, and bottom row corresponding to datasets P1-P2 and E1-E2. The dashed line corresponds to a ordinary least squares fit and dotted line corresponds to a weighted least squares fit.

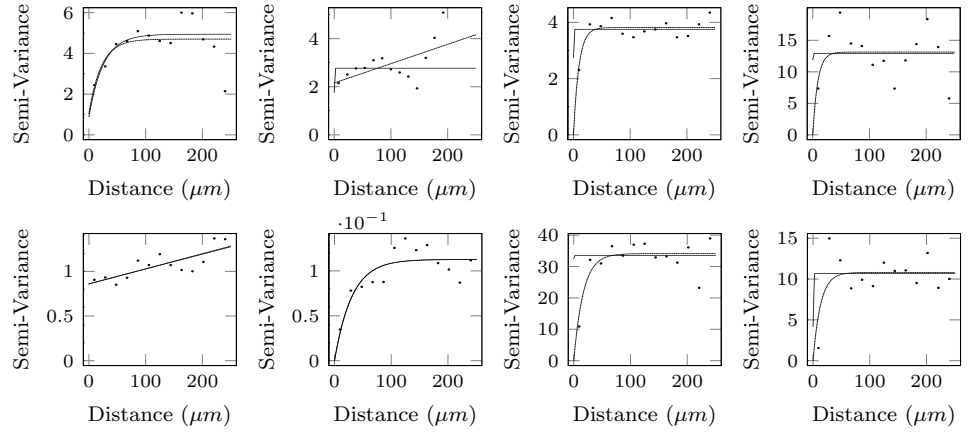


Figure C.10: Variograms for spatial Feature 3 with top row corresponding to datasets A1-A4, and bottom row corresponding to datasets P1-P2 and E1-E2. The dashed line corresponds to a ordinary least squares fit and dotted line corresponds to a weighted least squares fit.

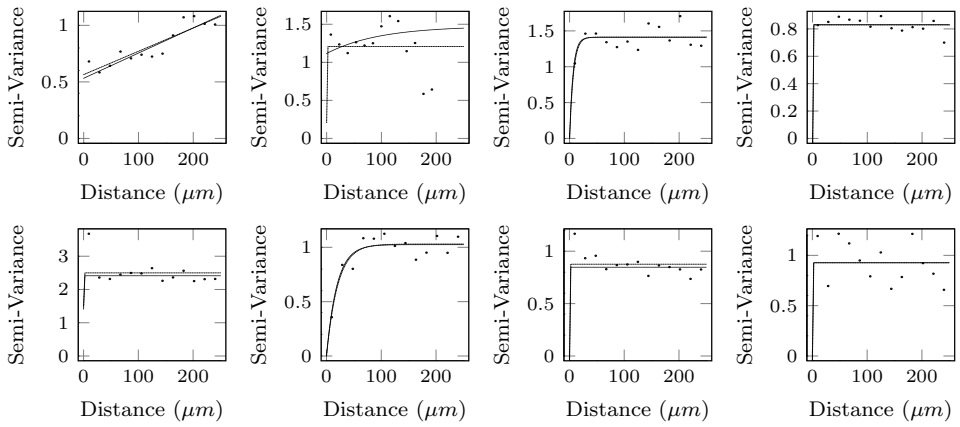


Figure C.11: Variograms for spatial Feature 4 with top row corresponding to datasets A1-A4, and bottom row corresponding to datasets P1-P2 and E1-E2. The dashed line corresponds to a ordinary least squares fit and dotted line corresponds to a weighted least squares fit.

REFERENCES

Amrein, M. and Künsch, H. R. (2012). Rate estimation in partially observed Markov jump processes with measurement errors. *Statistics and Computing*, 22(2):513–526.

Ananthasubramaniam, B., Herzog, E. D., and Herzel, H. (2014). Timing of neuropeptide coupling determines synchrony and entrainment in the mammalian circadian clock. *PLoS Computational Biology*, 10(4):e1003565.

Anderson, D. F. and Kurtz, T. G. (2011). Continuous time Markov chain models for chemical reaction networks. In *Design and Analysis of Biomolecular Circuits*, pages 3–42. Springer.

Andrieu, C., Doucet, A., and Holenstein, R. (2009). Particle Markov chain Monte Carlo for efficient numerical simulation. In *Monte Carlo and quasi-Monte Carlo methods 2008*, pages 45–60. Springer.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697–725.

Baddeley, A. and Gill, R. D. (1997). Kaplan-Meier estimators of distance distributions for spatial point processes. *The Annals of Statistics*, pages 263–292.

Baddeley, A., Rubak, E., Møller, J., et al. (2011). Score, pseudo-score and residual diagnostics for spatial point process models. *Statistical Science*, 26(4):613–646.

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, 42(3):283–322.

Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42. URL: `www.jstatsoft.org`, ISSN: 1548-7660.

Baddeley, A., Turner, R., Mateu, J., and Bevan, A. (2013). Hybrids of Gibbs point process models and their implementation. *Journal of Statistical Software*, 55(11):1–43.

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666.

Baddeley, A. J. and Van Lieshout, M. (1995). Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, 47(4):601–619.

Barbour, A. D. (1974). On a functional central limit theorem for Markov population processes. *Advances in Applied Probability*, pages 21–39.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Berman, M. and Diggle, P. (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 81–92.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, pages 179–195.

Besag, J. (1977). Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute*, 47(2):77–92.

Blake, W. J., Balázsi, G., Kohanski, M. A., Isaacs, F. J., Murphy, K. F., Kuang, Y., Cantor, C. R., Walt, D. R., and Collins, J. J. (2006). Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular Cell*, 24(6):853–865.

Bonnefont, X., Lacampagne, A., Sanchez-Hormigo, A., Fino, E., Creff, A., Mathieu, M., Smallwood, S., Carmignac, D., Fontanaud, P., Travo, P., et al. (2005). Revealing the large-scale network organization of growth hormone-secreting cells. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16880.

Boys, R. J. and Giles, P. R. (2007). Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates. *Journal of Mathematical Biology*, 55(2):223–247.

Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135.

Chabot, J. R., Pedraza, J. M., Luitel, P., and van Oudenaarden, A. (2007). Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock. *Nature*, 450(7173):1249–1252.

Chesson, P. (1978). Predator-prey theory and variability. *Annual Review of Ecology and Systematics*, 9:323–347.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

Choi, B. and Rempala, G. A. (2012). Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics*, 13(1):153–165.

Christian, H. C., Chapman, L. P., and Morris, J. F. (2007). Thyrotrophin-Releasing Hormone, Vasoactive Intestinal Peptide, Prolactin-Releasing Peptide and Dopamine Regulation of Prolactin Secretion By Different Lactotroph Morphological Subtypes in the Rat. *Journal of Neuroendocrinology*, 19(8):605–613.

Daigle, B. J., Roh, M. K., Petzold, L. R., and Niemi, J. (2012). Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinformatics*, 13(1):68.

Diggle, P. (1986). Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern. *Journal of Neuroscience Methods*, 18(1-2):115–125.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer Science & Business Media.

Diggle, P. J. (2005). Spatio-temporal point processes: Methods and Applications. In *Case Studies in Spatial Point Process Modeling*. Springer.

Dobrescu, R. and Purcarea, V. (2009). Network based models for biological applications. *Journal of Medicine and Life*, 2(2):176.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science Signaling*, 297(5584):1183.

Fearnhead, P., Giagos, V., and Sherlock, C. (2014). Inference for reaction networks using the Linear Noise Approximation. *Biometrics*, 70(2):457–466.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Featherstone, K., Harper, C. V., McNamara, A., Semprini, S., Spiller, D. G., McNeilly, J., McNeilly, A. S., Mullins, J. J., White, M. R. H., and Davis, J. R. E. (2011). Pulsatile patterns of pituitary hormone gene expression change during development. *Journal of Cell Science*, 124(20):3484–3491.

Ferm, L., Lötstedt, P., and Hellander, A. (2008). A hierarchy of approximations of the master equation scaled by a size parameter. *Journal of Scientific Computing*, 34(2):127–151.

Finkenstädt, B., Heron, E. A., Komorowski, M., Edwards, K., Tang, S., Harper, C. V., Davis, J. R. E., White, M. R. H., Millar, A. J., and Rand, D. A. (2008). Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24):2901–2907.

Finkenstädt, B., Woodcock, D. J., Komorowski, M., Harper, C. V., Davis, J. R. E., White, M. R. H., and Rand, D. A. (2013). Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: An application to single cell data. *The Annals of Applied Statistics*, 7(4):1960–1982.

Fleischer, F., Beil, M., Kazda, M., and Schmidt, V. (2006). Analysis of spatial point patterns in microscopic and macroscopic biological image data. In *Case Studies in Spatial Point Process Modeling*, pages 235–260. Springer.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*, volume 68. CRC Press.

Gardiner, C. W. (1985). *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer Berlin.

Gelman, A., Hwang, J., and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, pages 1–20.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. *Stochastic Geometry: Likelihood and Computation*, 80:79–140.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Gillespie, D. T. (2000). The chemical Langevin equation. *The Journal of Chemical Physics*, 113:297.

Golightly, A., Henderson, D. A., and Sherlock, C. (2014). Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing*, pages 1–17.

Golightly, A. and Wilkinson, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788.

Golightly, A. and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820.

Gomez-Uribe, C. A. and Verghese, G. C. (2007). Mass fluctuation kinetics: Capturing stochastic effects in systems of chemical reactions through coupled mean-variance computations. *The Journal of Chemical Physics*, 126(2):024109.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Grima, R. (2010). An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *The Journal of Chemical Physics*, 133(3):035101.

Groot, R. and Warren, P. (1997). Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *The Journal of Chemical Physics*, 107:4423.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Harper, C. V., Featherstone, K., Semprini, S., Friedrichsen, S., McNeilly, J., Paszek, P., Spiller, D. G., McNeilly, A. S., Mullins, J. J., and Davis, J. R. E. (2010). Dynamic organisation of prolactin gene expression in living pituitary tissue. *Journal of Cell Science*, 123(3):424.

Harper, C. V., Finkenstädt, B., Woodcock, D. J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D. G., Mullins, J. J., Rand, D. A., and Davis, J. R. E. (2011). Dynamic analysis of stochastic transcription cycles. *PLoS Biology*, 9(4):e1000607.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97.

Heron, E. A., Finkenstädt, B., and Rand, D. A. (2007). Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics*, 23(19):2596.

Hey, K., Momiji, H., Featherstone, K., Davis, J., White, M., Rand, D., and Finkenstädt, B. (2015). Inference for a transcriptional stochastic switch model from single cell imaging data. *Biostatistics*. DOI: 10.1093/biostatistics/kxv010.

Hobolth, A. and Stone, E. A. (2009). Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *The Annals of Applied Statistics*, 3(3):1204.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, volume 70. John Wiley & Sons.

Illian, J. B., Møller, J., and Waagepetersen, R. P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, 16(3):389–405.

Jenkins, D. J., Finkenstädt, B., and Rand, D. A. (2013). A temporal switch model for estimating transcriptional activity in gene expression. *Bioinformatics*, 29(9):1158–1165.

Jensen, J. L. and Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *The Annals of Applied Probability*, pages 445–461.

Kærn, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45.

Kazeev, V., Khammash, M., Nip, M., and Schwab, C. (2014). Direct solution of the chemical master equation using quantized tensor trains. *PLoS Computational Biology*, 10(3):e1003359.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25.

Komorowski, M., Finkenstädt, B., Harper, C., and Rand, D. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10(1):343.

Komorowski, M., Finkenstädt, B., and Rand, D. (2010). Using a single fluorescent reporter gene to infer half-life of extrinsic noise and other parameters of gene expression. *Biophysical Journal*, 98(12):2759–2769.

Kurtz, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58.

Kurtz, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356.

Kurtz, T. G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and Their Applications*, 6(3):223–240.

Larson, D. R., Singer, R. H., and Zenklusen, D. (2009). A single molecule view of gene expression. *Trends In Cell Biology*, 19(11):630–637.

Le Tissier, P., Hodson, D., Lafont, C., Fontanaud, P., Schaeffer, M., and Mollard, P. (2012). Anterior pituitary cell networks. *Frontiers in Neuroendocrinology*, 33(3):252–266.

Li, C.-W. and Chen, B.-S. (2009). Stochastic spatio-temporal dynamic model for gene/protein interaction network in early Drosophila development. *Gene Regulation and Systems Biology*, 3:191.

Lieshout, M. v. and Baddeley, A. (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361.

Lillacci, G. and Khammash, M. (2013). The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18):2311–2319.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087.

Mollard, P., Hodson, D. J., Lafont, C., Rizzoti, K., and Drouin, J. (2012). A tridimensional view of pituitary development and function. *Trends in Endocrinology & Metabolism*, 23(6):261–269.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. CRC Press.

Munsky, B. and Khammash, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4):044104.

Munsky, B. and Khammash, M. (2008). The finite state projection approach for the analysis of stochastic noise in gene networks. *Automatic Control, IEEE Transactions on*, 53(Special Issue):201–214.

Nelson, D. E., Ihekwaba, A. E. C., Elliott, M., Johnson, J. R., Gibney, C. A., Foreman, B. E., Nelson, G., See, V., Horton, C. A., and Spiller, D. G. (2004). Oscillations in NF-κ B signaling control the dynamics of gene expression. *Science Signaling*, 306(5696):704.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:3–48.

Ohser, J. (1983). On estimators for the reduced second moment measure of point processes. *Statistics: A Journal of Theoretical and Applied Statistics*, 14(1):63–71.

Opper, M. and Sanguinetti, G. (2010). Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, 26(13):1623–1629.

Owen, J., Wilkinson, D. J., and Gillespie, C. S. (In Press). Likelihood free inference for Markov processes: A comparison. *Statistical Applications in Genetics and Molecular Biology*.

Paszek, P., Ryan, S., Ashall, L., Sillitoe, K., Harper, C. V., Spiller, D. G., Rand, D. A., and White, M. R. (2010). Population robustness arising from cellular heterogeneity. *Proceedings of the National Academy of Sciences*, 107(25):11644–11649.

Paulauskas, N., Pranevicius, M., Pranevicius, H., and Bukauskas, F. F. (2009). A stochastic four-state model of contingent gating of gap junction channels containing two "fast" gates sensitive to transjunctional voltage. *Biophysical Journal*, 96(10):3936–3948.

Paulsson, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175.

Peccoud, J. and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222–234.

Penttinen, A. (1984). *Modelling interaction in spatial point patterns: Parameter estimation by the maximum likelihood method*, volume 7 of *Jyväskylä Studies in Computer Science, Economics, and Statistics*. Jyväskylän yliopisto.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models*. Springer.

Pivkin, I., Richardson, P., and Karniadakis, G. (2009). Effect of red blood cells on platelet aggregation. *Engineering in Medicine and Biology Magazine, IEEE*, 28(2):32–37.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raj, A. and Van Oudenaarden, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–226.

Rao, C. V. and Arkin, A. P. (2003). Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *The Journal of Chemical Physics*, 118(11):4999–5010.

Ribeiro Jr, P. J. and Diggle, P. J. (2001). geoR: A package for geostatistical analysis. *R News*, 1(2):14–18.

Ripley, B. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266.

Ripley, B. D. (1991). *Statistical Inference for Spatial Processes*. Cambridge University Press.

Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science Signaling*, 307(5717):1962.

Rubinstein, R. Y. (1997). Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112.

Ruttor, A. and Opper, M. (2009). Efficient statistical inference for stochastic reaction processes. *Physical Review Letters*, 103(23):230601.

Sanchez, A., Choubey, S., and Kondev, J. (2013). Stochastic models of transcription: From single molecules to single cells. *Methods*, 62(1):13–25.

Sanguinetti, G., Ruttor, A., Opper, M., and Archambeau, C. (2009). Switching regulatory models of cellular stress response. *Bioinformatics*, 25(10):1280–1286.

Schnoerr, D., Sanguinetti, G., and Grima, R. (2014). Validity conditions for moment closure approximations in stochastic chemical kinetics. *The Journal of Chemical Physics*, 141(8):084103.

Semprini, S., Friedrichsen, S., Harper, C. V., McNeilly, J. R., Adamson, A. D., Spiller, D. G., Kotelevtseva, N., Brooker, G., Brownstein, D. G., and McNeilly, A. S. (2009). Real-time visualization of human prolactin alternate promoter usage in vivo using a double-transgenic rat model. *Molecular Endocrinology*, 23(4):529–538.

Shapiro, B., Jönsson, H., Sahlin, P., Heisler, M., Roeder, A., Bulr, M., Meyerowitz, E., and Mjolsness, E. (2011). Tessellations and Pattern Formation in Plant Growth and Development. In van de Weygaert R, Vegter G, R. J. I. V., editor, *In Tessellations in the Sciences: Virtues, Techniques and Applications of Geometric Tilings*. Springer-Verlag. In Press.

Shapiro, B. and Mjolsness, E. (2001). Developmental simulations with Cellerator. In *Proceedings of the Second International Conference on Systems Biology*, pages 4–7.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Spiller, D. G., Wood, C. D., Rand, D. A., and White, M. R. (2010). Measurement of single-cell dynamics. *Nature*, 465(7299):736–745.

Stathopoulos, V. and Girolami, M. A. (2013). Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984).

Stein, A. and Georgiadis, N. (2006). Spatial marked point patterns for herd dispersion in a savanna wildlife herbivore community in Kenya. In *Case Studies in Spatial Point Process Modeling*, pages 261–273. Springer.

Stephens, D. J. and Allan, V. J. (2003). Light microscopy techniques for live cell imaging. *Science Signaling*, 300(5616):82.

Strauss, D. J. (1975). A model for clustering. *Biometrika*, 62(2):467–475.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472.

Symeonidis, V., Em Karniadakis, G., and Caswell, B. (2005). Dissipative particle dynamics simulations of polymer chains: Scaling laws and shearing response compared to DNA experiments. *Physical Review Letters*, 95(7):76001.

Terry, A. J., Sturrock, M., Dale, J. K., Maroto, M., and Chaplain, M. A. (2011). A spatio-temporal model of Notch signalling in the zebrafish segmentation clock: conditions for synchronised oscillatory dynamics. *PLoS One*, 6(2):e16980.

Thomas, P., Popović, N., and Grima, R. (2014). Phenotypic switching in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 111(19):6994–6999.

Thomas, P., Straube, A. V., and Grima, R. (2012). The slow-scale linear noise approximation: An accurate, reduced stochastic description of biochemical networks under timescale separation conditions. *BMC Systems Biology*, 6(1):39.

Tkačik, G. and Walczak, A. M. (2011). Information transmission in genetic regulatory networks: A review. *Journal of Physics: Condensed Matter*, 23(15):153102.

Turner, R. (2009). Point patterns of forest fire locations. *Environmental and Ecological Statistics*, 16(2):197–223.

Ullah, M. and Wolkenhauer, O. (2009). Investigating the two-moment characterisation of subcellular biochemical networks. *Journal of Theoretical Biology*, 260(3):340–352.

van Kampen, N. G. (1961). A power series expansion of the master equation. *Canadian Journal of Physics*, 39(4):551–567.

Vogel, R. and Weingart, R. (1998). Mathematical model of vertebrate gap junctions derived from electrical measurements on homotypic and heterotypic channels. *The Journal of Physiology*, 510(1):177–189.

Wallace, E. W. J., Gillespie, D. T., Sanft, K. R., and Petzold, L. R. (2012). Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *Systems Biology, IET*, 6(4):102–115.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 9999:3571–3594.

Widom, B. and Rowlinson, J. S. (1970). New model for the study of liquid–vapor phase transitions. *The Journal of Chemical Physics*, 52(4):1670–1684.

Wilkinson, D. J. (2011). *Stochastic Modelling for Systems Biology*, volume 44. CRC press.

Zechner, C., Unger, M., Pelet, S., Peter, M., and Koeppl, H. (2014). Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods*, 11(2):197–202.

ABBREVIATIONS

A1 Adult Dataset 1
A2 Adult Dataset 2
A3 Adult Dataset 3
A4 Adult Dataset 4
ABC Approximate Bayesian Computation
AIC Akaike's Information Criteria
BDA Birth Death Approximation
BDD Birth Death Decomposition
CLE Chemical Langevin Equation
CSR Complete Spatial Randomness
DA Diffusion Approximation
DIC Deviance Information Criteria
E1 Embryonic Day 18.5 Dataset 1
E18.5 Embryonic Day 18.5
E2 Embryonic Day 18.5 Dataset 2
EM Expectation Maximisation
EMRE Effective Mesoscopic Rate Equation
FFSB Forward Filtering Backward Sampling
FSP Finite State Projection
GFP Green Fluorescent Protein
HMM Hidden Markov Model
LHS Left Hand Side
LNA Linear Noise Approximation
MA Moment Closure Approximation
MCEM Monte Carlo Expectation Maximisation
MCMC Markov Chain Monte Carlo
ME Master Equation
MH Metropolis Hastings

MJP Markov Jump Process
MLE Maximum Likelihood Estimate
mRNA Messenger RNA
MSE Mean Square Error
ODE Ordinary Differential Equation
P1 Post-natal Day 1.5 Dataset 1
P1.5 Post-natal Day 1.5
P2 Post-natal Day 1.5 Dataset 2
PDE Partial Differential Equation
PMMH Particle Marginal Metropolis Hastings
PRL Prolactin
QSSA Quasi Steady State Approximation
RHS Right Hand Side
RJ Reversible Jump
RRE Reaction Rate Equation
RW Random Walk
s.d. Standard Deviation
SDE Stochastic Differential Equation
SIR Sequential Importance Resampling
SIS Sequential Importance Sampling
SMC Sequential Monte Carlo
SRN Stochastic Reaction Network
SSA Stochastic Simulation Algorithm
ssLNA slow-scale LNA
SSM Stochastic Switch Model
TT Translation Transformed
WAIC Widely Applicable Information Criteria